2018-2019

Master 1 Sciences de l'Information et des Bibliothèques

# L'archivage du Web français par la Bibliothèque nationale de France

## Une nouvelle approche des missions de la BnF ?

**Zielinski Océane**

Sous la direction de Mme Alibert Florence

Membres du jury
Alibert Florence | Directrice de mémoire, Directrice du M2 Sciences de l'Information et des Bibliothèques
Neveu Valérie | Directrice du M1 Sciences de l'Information et des Bibliothèques



Soutenu le 18 juin 2019

FACULTÉ DES LETTRES, LANGUES ET SCIENCES HUMAINES
UNIVERSITÉ D'ANGERS

# REMERCIEMENTS

Je tiens à remercier Madame Alibert, ma directrice de mémoire, qui m'a aidée et conseillée tout au long de la rédaction de cette étude et m'a faite découvrir l'archivage du Web.

Je remercie également les membres du dépôt légal numérique de la BnF pour l'aide précieuse qu'ils m'ont apporté et le temps qu'ils m'ont accordé, notamment pour la réalisation d'un entretien.

Enfin, je remercie mes correctrices, ma mère Ingrid, ma sœur Gwenäelle et Margot et Myriam, pour leurs relectures assidues et leurs conseils.

# Sommaire

# Introduction

*« Archiver le Web, c'est être condamné aux vides[1]. »*

L'archivage du Web est un paradoxe en soi. Lorsqu'au milieu des années 1990, les premières archives du Web sont apparues, elles répondaient déjà au paradoxe créé par Internet. Jamais il n'y avait eu autant d'informations circulant si librement et en telle quantité. Jamais il n'y avait eu autant de perte d'informations aussi massive et rapide, en dehors d'un contexte de conflit ou de censure. Internet a permis à tout un chacun de publier ses créations, de les diffuser au-delà des frontières, mais rien n'en assure la diffusion et l'accessibilité sur le long terme. En moyenne, une page Web a une durée de vie de 100 jours avant d'être modifiée ou de disparaître[2]. La nécessité de conserver cette connaissance, ce patrimoine en péril, s'est faite ressentir très tôt, quelques années à peine après l'ouverture d'Internet au grand public, en 1990. On archive donc le Web pour le conserver, mais en faisant cela, on n'en n'obtient qu'un instantané. Pour diverses raisons, les archives du Web proposent au mieux le portrait d'une partie du Web à un instant T, jamais dans sa totalité ni sa parfaite intégrité. Les archives du Web sont constituées de trou, de vide que l'on comble au mieux, pour proposer le portrait inanimé d'un Web toujours en mouvement, un flux continu où même la notion de temporalité se trouble. Les archives du Web constituent la seule trace des productions mises en ligne et qui ont pu disparaître. Il s'agit aussi d'un reflet de nos sociétés au fil des captures. Ce travail titanesque est donc nécessaire pour éviter la perte d'une partie du patrimoine mondial, en perpétuelle expansion. Il est cependant illusoire de penser que l'on peut archiver l'intégralité du Web mondial. C'est pourquoi les initiatives se sont multipliées pour assurer une meilleure couverture du Web.

L'archivage du Web commence à être assez bien documenté. Les professionnels impliqués dans ces projets ont beaucoup publié sur leurs méthodologies de travail, les différentes méthodes existantes, les obstacles rencontrés, les évolutions apportées sur les procédés techniques. Les chercheurs, de leur côté, ont mis l'accent sur l'utilisation de ces archives du Web, sur les outils disponibles et nécessaires et sur l'importance croissante

---

[1] DE LA PORTE Xavier, "Archiver le Web, c'est être condamné aux vides", *La vie numérique*, France Culture, 2016. https://www.franceculture.fr/emissions/la-vie-numerique/archiver-le-web-cest-etre-condamne-aux-vides

[2] DE LA PORTE Xavier, « Brewster Kahle, Internet Archive : « le meilleur du web est déjà perdu » », *internetactu.net*, 2011. http://www.internetactu.net/2011/06/28/brewster-kahle-internet-archive-le-meilleur-du-web-est-deja-perdu/

qu'elles vont représenter pour la recherche dans les décennies à venir. Malgré cet intérêt nouveau pour ces archives, un pan reste assez peu exploré : l'aspect bibliothéconomique. Gildas Illien, ancien chef du service du dépôt légal numérique à la BnF, a beaucoup écrit entre 2004 et 2011 sur ce point. Dans l'optique de présenter et d'expliquer au mieux ce qu'est l'archivage du Web et le fonctionnement de son service, il s'est intéressé aux évolutions à venir pour les métiers des bibliothèques en lien avec cette nouvelle mission et surtout, ce nouveau type d'archive intégralement dématérialisée. Depuis, l'étude des répercussions bibliothéconomiques de l'archivage du Web à la BnF et dans d'autres bibliothèques a été délaissée. C'est donc sous cet angle que nous avons souhaité aborder l'archivage du Web en France. Nous nous sommes posés la question en quoi l'archivage du Web a-t-il apporté une nouvelle approche des missions de la Bibliothèque nationale de France.

Pour y apporter des éléments de réponse, cette étude a été divisée en deux temps. D'abord, il est nécessaire de retracer les grandes lignes de l'archivage du Web dans le monde. Il y sera question de la naissance de l'archivage du Web et des problématiques qui l'ont entouré dès le début. Une présentation non-exhaustive des programmes d'archivage à travers le monde puis en France permettra d'aborder les différentes pratiques.

Ensuite, nous consacrerons une étude de cas à la BnF et ses systèmes de collecte. Il s'agira d'étudier les différentes procédures de collecte en comparant la réalisation de deux collectes d'urgences autour des attentats de Paris en 2015 avec les collectes programmées. Enfin, nous nous intéresserons aux évolutions des métiers des bibliothèques apportés par l'archivage du Web.

# Historique de l'archivage dans le monde et en France

## 1. Naissance et problématiques de l'archivage du Web

La fin des années 1990 marque le début de l'archivage du Web à travers le monde. D'abord par des informaticiens un peu visionnaires et idéalistes, qui ont très vite été rejoints par des bibliothécaires partageant leur engouement pour la préservation du numérique. Forte de cet intérêt nouveau pour le document numérique et sa conservation, une réflexion autour de l'encadrement juridique de ces documents et de leur archivage s'est développée dans bon nombre de pays.

### 1.1. 1996 : la prise de conscience

L'année 1996 marque l'an un de l'archivage du web. Les premières prises de conscience sur l'importance et l'ampleur que prend le contenu disponible sur le Web commencent à percer. Quelques personnes décident d'agir face à la perte en cours, et à venir, de ces informations en créant les premières initiatives d'archivage du Web. À l'origine de ce projet, on trouve Brewster Kahle.

#### 1.1.1. Brewster Kahle et Internet Archive

Dès 1995, l'idée d'archiver le Web mondial germe dans l'esprit de Brewster Kahle, ingénieur informaticien issu du MIT. Très sensible au problème des erreurs 404 de plus en plus fréquentes, seules traces restantes d'un site ou d'une page web disparus, Brewster Kahle a réfléchi à un moyen d'éviter cette perte. Le support numérique est fragile, ce qui est encore plus vrai pour le Web où la durée de vie d'une page n'est que de quelques mois. Il souhaitait aussi éviter que l'on reproduise l'erreur de la destruction de films en 1997 pour récupérer l'argent contenu dans les pellicules[3]. L'archivage de ces pages avant leur disparition lui est donc apparu comme la solution la plus évidente et la plus efficace pour la conservation. Internet commençait à être vu comme une bibliothèque d'un nouveau genre au début des années 1990 et c'est dans cette optique que Kahle a créé la fondation Internet Archive. Officiellement lancée en avril 1996 à San Francisco, Internet Archive a pour but de collecter et

[3] MUSIANI Francesca, SCHAFER Valérie, « Patrimoine et patrimonialisation numériques », in *Reset*, 2017, n° 6, p. 1.

archiver tout ce que l'humanité pouvait publier et créer sur internet afin que chacun puisse bénéficier d'un accès à cette connaissance. Cela impliquait donc de collecter les sites web, mais aussi les livres, la musique et les vidéos[4].

## Créations d'outils et méthodes adaptés

La méthode appliquée par Internet Archive est assez simple. Tout ce qui est publié sur Internet est considéré comme public donc peut être collecté. C'est une approche intégrale et quelque peu agressive. En effet, on collecte absolument tout, sans opérer de sélection préalable et sans se soucier du droit d'auteur. C'est la loi américaine du *opt-out* qui s'applique ici, c'est donc au propriétaire d'un contenu de se manifester pour le faire retirer de l'accès public des archives. En 1996, les collectes tests se font manuellement et servent essentiellement à expérimenter le modèle de collecte. Une première expérimentation pour un tiers a eu lieu cette même année avec la Smithsonian Institution à Washington DC. Internet Archive a collecté pour elle tous les sites de candidats à la présidentielle des États-Unis. Cette campagne test a permis de développer des outils pour capturer ces sites dans leur intégralité[5].

À la fin de l'année 1997, Brewster Kahle a créé un outil commercial permettant de butiner, collecter et indexer les pages web : Alexa. Conçu comme un *plug-in* qui collecte également les informations sur la fréquentation des sites, leur fréquence de renouvellement, le nombre de liens rattachés, etc., le *crawler* Alexa permet également d'accéder aux versions antérieures des sites en allant directement les chercher dans les archives d'Internet Archive. Cet outil constitue le premier accès public aux collections de la fondation[6]. Cette même année, Alexa était programmée pour exécuter une capture du Web toutes les huit semaines. Le Web était encore assez limité dans sa taille, rendant l'opération possible. Cependant, Internet Archive a rencontré deux difficultés principales lors de cette première année d'archivage. D'une part, l'accès aux pages web était lent, même très lent, pouvant aller de quelques secondes à quelques jours selon le matériel et la disponibilité de la bande passante. D'autre part, la collecte intégrale et l'utilisation du système *opt-out* ont été critiquées. Un protocole d'exclusion des robots a été développé, il suffisait d'intégrer le fichier robots.txt au code source du site pour que le *crawler* passe son chemin sans rien collecter. Force est de constater que les

[4] MUSIANI Francesca, *et al.*, *Qu'est-ce qu'une archive du Web ?*, Marseille, Open Édition Press, 2019, p. 57.

[5] MASANÈS Julien (dir.), *Web archiving*, Berlin, Springer, 2006, p. 202.

[6] CHAIMBAULT Thomas, *L'archivage du web*, Villeurbanne, Enssib, 2008, p. 29-30.

protestations envers le système *opt-out* n'ont pas porté leurs fruits, car il est encore aujourd'hui utilisé par Internet Archive[7].

Alexa a permis de passer à un modèle de collecte automatisé jusque dans l'indexation, ce qui représente des économies de temps mais aussi d'argent, car la collecte mobilise moins de personnes. Le Web n'a cessé de voir son volume augmenter dès 1998, rendant de plus en plus difficile une collecte manuelle exhaustive. Le nombre de pages doublait tous les trois à six mois, tout comme le nombre d'utilisateurs. Pour suivre cette augmentation du nombre de pages à collecter, Alexa s'est associée à deux acteurs majeurs de l'informatique et du Web : Microsoft et Netscape. Forte de ce partenariat, Alexa a pu intégrer son *plug-in* aux moteurs de recherche Internet Explorer et Netscape, couvrant ainsi près de 90 % des postes informatiques du globe[8]. Brewster Kahle a vendu Alexa à Amazon en 1999.

## Le temps des projets

Entre 2000 et 2003, Internet Archive s'est lancé dans plusieurs projets d'envergure. D'abord, le format ARC, créé en 1996 par Mike Burner et Brewster Kahle est devenu le format fixe pour l'archivage du web et s'impose rapidement comme la norme mondiale. Forte du succès de sa campagne test avec la Smithsonian Institution, Internet Archive est contactée par la Bibliothèque du Congrès pour réaliser une collecte autour des élections de 2000 puis de 2002. Cette mission marque le passage d'Internet Archive comme projet expérimental à une institution reconnue et solidement établie[9], amenant par la même occasion une croissance considérable de ses archives. Ces dernières ont triplé entre mars 2000 et mars 2001, avec une augmentation d'environ 40 téraoctets par mois. En septembre 2001, Internet Archive se trouve confronté à sa première collecte « d'urgence » avec les attentats du 11 septembre. Entre le 11 septembre et le 1er décembre, elle collecte en association avec la Bibliothèque du Congrès, des images venant de plus de 30 000 sites ainsi que des centaines d'heures d'émissions de presse. Cette collecte a été une sorte d'épreuve du feu pour tester l'efficacité et les capacités des outils de collecte face au besoin de réactivité[10].

En parallèle, une solution pour permettre facilement un accès aux collections pour le public était recherchée. Bien qu'une partie des archives soient accessibles grâce au *plug-in* Alexa, le reste des collections nécessitait d'avoir quelques connaissances en programmation

[7] MASANES Julien, op. cit., p. 203-204.

[8] Ibid. p. 205.

[9] Ibid. p. 203, 206.

[10] Ibid. p. 207-208.

pour y accéder. Le 24 octobre 2001, la solution est officiellement dévoilée au public avec la *Wayback Machine*, sorte de machine à remonter le temps d'Internet. Créée par Alexa Internet, sous contrat avec Internet Archive, la *Wayback Machine* donne accès aux archives via les URL des sites, permettant l'accès à quelques 10 milliards de pages web archivées jusqu'alors, représentant 100 To de données[11]. Cet outil s'est d'ailleurs répandu dans bon nombre d'initiatives d'archivage du Web, notamment à la BnF.

Toujours dans l'optique de créer une bibliothèque numérique mondiale ouverte à tous, Internet Archive s'est lancée dans quatre projets majeurs en 2002. Il a d'abord été question de créer un site miroir de ses collections, une sorte d'archive des archives. Parfaitement consciente de la fragilité du support physique des serveurs et *data centers*, risque accentué par la présence d'une faille sismique majeure sous la ville de San Francisco, Internet Archive a cherché un endroit où implanter des serveurs miroirs. C'est la ville d'Alexandrie en Égypte qui a été choisie, alors que le projet de la construction d'une nouvelle grande bibliothèque était en cours. Internet Archive a ainsi envoyé des serveurs et plus de 100 To de données sur place. Tout était installé et opérationnel pour l'ouverture de la bibliothèque en avril 2002. Ce site miroir partage l'intégralité des archives d'Internet Archive à San Francisco, quasiment en temps réel. L'idée est notamment de créer une bibliothèque d'Alexandrie 2.0 tout en évitant de reproduire la catastrophe antique en multipliant les lieux de stockage. Un autre site miroir a, par la suite, été implanté à Amsterdam[12].

Ensuite, pour aller plus loin dans l'idée de rendre accessible à tous et partout le contenu du Web, la fondation a créé l'*Internet Bookmobile*. Il s'agit d'un van équipé pour pouvoir consulter les archives du Web, télécharger et imprimer un million de livres numérisés, le tout gratuitement. Une première *Internet Bookmobile* a été lancée aux États-Unis puis rapidement une autre a vu le jour en Inde, puis au Kenya. Si le projet n'a pas pu perdurer longtemps sur le sol américain après sa tournée contre l'extension des copyrights, sans succès, le projet s'est relativement bien développé à l'étranger avec la création de dix autres vans rien qu'en Inde[13].

Ensuite, Internet Archive s'est lancée dans la création de deux collections d'envergure. Juin 2002 vit ainsi la naissance de la première collection de livres d'Internet Archive. Août de la même année a vu le lancement de la collection musique qui inclut également la *Live Music Archive* regroupant des concerts téléchargeables gratuitement[14].

---

[11] Ibid. p. 207.

[12] DE LA PORTE Xavier, « Brewster Kahle, Internet Archive : « le meilleur du web est déjà perdu » ». ; MASANES Julien, op. cit., p. 208-209.

[13] Ibid. p. 209.

[14] Ibid.

Enfin, Internet Archive s'est engagée en faveur de l'accès à un contenu de qualité à destination des enfants en leur donnant un accès global et direct via internet. L'*International Children's Digital Library* a ainsi été créée en partenariat avec l'Université du Maryland, avec le soutien de la Bibliothèque du Congrès et aussi d'acteurs majeurs du numérique comme Adobe Systems Inc[15].

S'il est difficile d'avoir des données chiffrées précises sur le volume collecté par Internet Archive à ce jour, le site internet donne une approximation de l'envergure de ses collections. Après un peu plus de 20 ans d'archivage, on peut accéder à quelques 330 milliards de pages web archivées, 20 millions de livres et textes, 4,5 millions d'enregistrements audio incluant 180 000 concerts, 4 millions de vidéos incluant 1,6 millions d'émissions et journaux télévisés, 3 millions d'images et 200 000 logiciels et programmes[16].

### 1.1.2. Les initiatives de la première heure

Internet Archive a eu très tôt une résonance à travers le monde. Quelques mois après son lancement, la Conférence des directeurs de bibliothèques nationales (CDNL) s'emparait du sujet de la sauvegarde des documents nativement numériques et du Web. Elle a alerté ses membres de l'importance d'archiver et de conserver ces documents, car ils représenteront bientôt un pan entier de la recherche et de la production de savoir, mais aussi la seule trace d'un passé numérique[17]. Deux pays prennent des mesures en 1996 pour archiver leur Web national, devenant des pionniers de l'archivage du Web aux côtés d'Internet Archive.

#### PANDORA

En juin 1996, la Bibliothèque nationale d'Australie, en partenariat avec les archives nationales, a lancé le projet PANDORA, *Preserving and Accessing Networked Documentary Ressources of Australia*. Jusqu'à la fin de l'année 1997, l'équipe s'est concentrée sur le développement d'une politique documentaire et la définition des procédures de sélection, de collecte et d'archivage qui seraient employées. Ce travail aboutit à la réalisation d'une première collecte à la fin de l'année en archivant 229 documents[18]. PANDORA fonctionne sur le

---

[15] Ibid.

[16] https://archive.org/about/

[17] CHAIMBAULT Thomas, op. cit., p. 25.

[18] GHARSALLAH Mehdi, « Archivage du Web français et dépôt légal des publications électroniques », *Documentalistes - Sciences de l'Information*, ADBS, 2004, p. 6.

principe d'une approche sélective semi-automatisée, contrairement à l'approche intégrale automatisée d'Internet Archive. Avec l'aide de dix partenaires, une sélection des sites à capturer est établie en amont. Cette sélection se base sur des critères stricts de qualité et de pertinence du contenu. Les sites retenus sont ensuite collectés à intervalles réguliers[19].

#### KulturarW³

La Suède s'est elle aussi penchée sur l'archivage du Web en septembre 1996. La Bibliothèque Royale lance alors son projet KulturarW³, jeu de mots entre *kulturarv* qui désigne le patrimoine culturel en suédois et le www de World Wide Web. Ici, c'est une approche exhaustive qui a été choisie pour collecter l'intégralité du web national, c'est-à-dire tous les sites avec un domaine en .se, sans critère de qualité ou pertinence. Ce choix d'exhaustivité a été justifié par la difficulté de prévoir quel contenu sera important pour la recherche ou les archives dans 20 ou 80 ans. Ces balayages réguliers par des *crawler* permettent d'archiver le Web national à moindre frais, cependant il n'y pas eu de véritable indexation au départ, ce qui complique l'utilisation de ces archives. Jusqu'en 2003, ce projet faisait partie de la NWA, *Nordic Web Archive*, qui regroupait les initiatives scandinaves. En 2003, ce projet a disparu en rejoignant l'IIPC, l'*International Internet Preservation Consortium*[20].

Malgré l'appel de la CDNL et ces trois initiatives de la première heure, il y a eu peu de nouveaux projets créés entre 1996 et 2003. Seules six initiatives ont vu le jour durant cette période[21]. Mais l'appel de l'Unesco à préserver le patrimoine nativement numérique a permis de redonner de l'élan à l'archivage du Web et à des bibliothèques qui regardaient de près ce qui se faisait chez Internet Archive, en Australie et en Suède.

## 1.2. Un patrimoine nativement numérique

L'essor de l'informatique et par extension l'essor du Web ont amené une importante production de documents et contenus créés dès leur origine sur un support numérique. Une prise de conscience de l'intérêt de cette production s'est opérée assez rapidement, à tel point que dès le début des années 2000, la communauté internationale s'empara du sujet. Il fallait définir ce nouveau patrimoine pour mieux l'appréhender et envisager sa conservation.

---

[19] CHAIMBAULT Thomas, op. cit., p. 26-31, 32 ; GHARSALLAH Mehdi, op. cit., p. 6.

[20] CHAIMBAULT Thomas, op. cit., p. 26, 30-31 ; MUSIANI Francesca, *et al.*, op. cit., p. 15.

[21] Ibid.

### 1.2.1 Un nouveau type de patrimoine

Lorsqu'en 2003 l'Unesco publie sa charte sur la conservation du patrimoine numérique, un nouveau type de patrimoine est défini. Jusqu'alors le patrimoine numérique désignait le patrimoine documentaire numérisé. Cette charte a introduit un nouveau terme, le *born-digital heritage*[22], que l'on traduit par « patrimoine nativement numérique ». Ce terme englobe ainsi l'ensemble des contenus, logiciels et formats créés directement sur support numérique. Avec cette démarche, l'Unesco répond au besoin de plus en plus pressant de diversifier la notion de patrimoine, qui ne concernait jusqu'en 2003 que le patrimoine physique, comme l'architecture, les livres ou manuscrits, etc. Cette charte pour la conservation du numérique s'insère dans un projet plus vaste de l'organisation, trouvant son origine dans le programme *Memory of the World,* débuté en 1992. L'Unesco a reconnu en 2003 l'existence du patrimoine culturel immatériel, dont fait partie le numérique[23].

Dans sa charte, l'Unesco donne des prérogatives pour assurer la conservation de ce patrimoine mondial en expansion permanente et exponentielle. D'abord, il est question de faire coopérer les différents acteurs du numérique et du patrimoine. Il est demandé à ces professionnels du public et du privé de s'unir afin de travailler à la conservation optimale de ce patrimoine, qui regroupe aussi bien des textes et des bases de données que des documents audio et/ou vidéo, des images fixes ou animées, des logiciels, les jeux-vidéo ou encore des pages web et les mails[24]. Ensuite, l'Unesco insiste sur le caractère prioritaire de la conservation de ce patrimoine nativement numérique. Puisqu'il est nouveau, à peine une décennie de mise en service publique et mondiale, il est nécessaire de s'intéresser aux moyens de le collecter et de le conserver. Déjà en 2003, le caractère fragile du support numérique avait été perçu[25]. Les supports de stockage sont éphémères si on les compare à la longévité d'un manuscrit. Un disque dur a une durée de vie moyenne de cinq ans avant de commencer à se démagnétiser. Or, comment conserver et sauvegarder ce patrimoine nativement numérique, par définition de plus en plus volumineux, si le support même de cette sauvegarde a une obsolescence aussi rapide ? C'est cette fragilité même du support qui donne un surcroît d'importance à ce patrimoine. On se retrouve dans un cas inédit d'urgence perpétuelle dans le choix de conservation. Les productions numériques étant éphémères et n'ayant que très peu de recul,

---

[22] Mentionné par deux fois aux articles 1er et 7e de la charte pour la conservation du patrimoine numérique.

[23] MUSIANI Francesca, *et al.*, op. cit., p. 15-16, 21.

[24] SCHAFER Valérie, MUSIANI Francesca, BORELLI Marguerite, « Le patrimoine culturel immatériel pour aider à penser le patrimoine activement numérique », *Patrimoine culturel immatériel et numérique*, M. SEVERO, S. CACHAT éd., Paris, L'Harmattan, 2017, p. 132.

[25] MUSIANI Francesca, *et al.*, op. cit. p. 21.

on privilégie d'abord une conservation plus massive, ne faisant pas une sélection trop restrictive au départ, pour avoir la meilleure représentativité de cette production pour les chercheurs et générations futurs. On leur laisse finalement le soin de définir eux-mêmes ce qui a un véritable intérêt patrimonial[26]. Ce faisant, il est essentiel de prendre en compte l'ensemble de l'objet nativement numérique que l'on souhaite conserver, à commencer par le support de consultation mais aussi tout son contexte de création et d'interactions.

### 1.2.2 Les moyens techniques disponibles

Le patrimoine numérique est en perpétuelle évolution aussi bien du point de vue du contenu que du point de vue technique. Contrairement à un livre qui garde son unité et sa compréhensibilité à travers le temps, ce n'est pas nécessairement le cas pour un objet nativement numérique. Ce dernier n'est pas un simple document unifié, homogène et qui conserve son statut d'un support à l'autre. Il faut s'assurer de la conservation de cet objet dans son intégralité, avec toutes les données, liens hypertextes, contextes, images, etc. qui le composent, ainsi que son support de consultation. Le support de médiation constitue la première difficulté et sans doute la première caractéristique du patrimoine nativement numérique. Matteo Treleani le dit très bien : « le document numérique ne peut [...] se passer de médiation, au point que, sans médiation, il ne peut tout simplement pas exister »[27]. On comprend donc qu'il est primordial de prendre en compte la pérennité et la conservation du support lorsqu'on archive un objet nativement numérique. C'est pourquoi dès le début de l'archivage et la conservation de ce patrimoine l'unité objet-support a été au cœur des travaux de recherche et d'expérimentation[28].

#### Cinq problèmes techniques majeurs

La conservation du numérique se heurte à plusieurs problèmes techniques. Howard Besser, spécialiste et pionnier dans le domaine de la préservation du numérique, en a identifié cinq. D'abord, l'installation et la maintenance des infrastructures pour la lecture des documents numériques archivés. Cela représente notamment un coût initial important en termes d'achat puis d'entretien. Les serveurs doivent par exemple être changés au plus tard tous les dix ans. Il faut également avoir les outils et connaissances pour décoder les logiciels

[26] CHAIMBAULT Thomas, op. cit., p.10.

[27] Cf. TRELEANI Matteo, *Qu'est-ce que le patrimoine numérique ? Une sémiologie de la circulation des archives*, Lormont, Le Bord de l'eau, 2017, p. 32.

[28] CHAIMBAULT Thomas, op. cit. p. 9-10, 18.

de compression ou de protection des pages web que l'on souhaite archiver. Comme nous l'avons déjà mentionné, l'intégrité du contexte du document doit aussi être conservé, ce qui inclut les liens hypertextes. En cela, la conservation et l'ajout de métadonnées détaillées sont essentiels afin de conserver toute la contextualisation du document original. Leur standardisation est même conçue au niveau international avec la *Dublin Core Metadata Initiative* (DCMI) qui travaille sur la norme Dublin Core. Ensuite, il faut définir un standard de bonne pratique ainsi qu'une politique d'acquisition. Cela permet de composer et délimiter des collections dans ces archives, tout en conservant les informations de provenance, mais aussi de garantir l'authenticité du document archivé. Enfin, il faut prendre en considération les problèmes de migration lors du transfert d'un support vers un autre, ce qui peut avoir des répercussions sur le document, dans son affichage ou son contenu[29].

Ces problèmes ne sont cependant pas restés sans solution. Pour pallier le risque de perte de données lors de la collecte, la solution la plus simple consiste à copier l'ensemble des données sans apporter de modification ou de sélection. On peut ainsi stocker une copie parfaite sur le même logiciel ou support. À titre d'exemple, c'est ce que l'on fait lorsque on copie une sauvegarde de ses fichiers sur un disque dur externe. Cela évite certes la perte de données mais ne résout pas le problème de la perte de lisibilité sur le long terme puisque le support en lui-même peut devenir obsolète. On peut également choisir de migrer un document vers un nouveau logiciel ou un nouveau matériel plus pérenne afin d'assurer une lisibilité plus durable. Dans le cas d'un support, on peut utiliser l'émulation qui permet de conserver l'intégrité du document numérique tout en conservant le contexte de lecture. Il s'agit d'une simulation du support. La BnF, par exemple, utilise l'émulation pour permettre de jouer aux jeux vidéo qu'elle collecte dans le cadre du dépôt légal[30]. Une solution hybride entre l'émulation et la migration existe également. On parle alors d'ordinateur virtuel universel qui reconstitue la forme d'origine d'un document ainsi que son environnement initial. Enfin, il est possible d'encapsuler les documents et leur support pour conserver leur unité[31]. On pourra ajouter un dernier problème, que Besser n'inclut pas, à savoir l'espace de stockage. Stocker tous ces documents, leurs supports et leurs métadonnées nécessite une capacité de stockage considérable. La plupart du temps, on ne se contente pas d'une unique sauvegarde de ces documents, le risque d'avarie matérielle étant trop grand. On conserve donc une copie de cette

---

[29] Ibid. p. 15. ; BESSER Howard, « Digital Longevity », in SITTS, Maxine, *Handbook for Digital Projects: A Management Tool for Preservation*, Andover, Mass. : Northeast Document Conservation Center, 2000, cité dans LYMAN, Peter, « Archiving the world wide web » [en ligne], in Council on library and Information resources, *Building a National Strategy for Preservation: Issues in Digital Media Archiving*. Berkeley, 2002.

[30] TRELEANI Matteo, op. cit. p. 35.

[31] CHAIMBAULT Thomas, op. cit. p. 19.

sauvegarde sur un autre serveur, un site miroir. Ce site miroir est stocké dans un autre lieu que le premier pour maximiser la sécurité des données. Avoir une copie conservée au même endroit ne servirait à rien en cas d'incendie dans la salle des serveurs par exemple. On profite souvent de ce site miroir pour y stocker des copies de conservation de très haute qualité, quand au contraire, on privilégie des sauvegardes plus « légères » pour les serveurs destinés à la mise en ligne publique. Toutefois, cette double conservation implique de doubler les capacités de stockage et donc de doubler le coût d'entretien et de fonctionnement du service. Il en va de même pour la charge de travail, car il faut du personnel et du temps pour mener les opérations de conservation préventives sur les deux sites[32].

### La gestion des risques

Aux problèmes énumérés par Howard Besser s'ajoutent quatre types de risques pour la conservation. Le premier regroupe les risques dits globaux, c'est-à-dire liés à l'environnement, à la sécurité informatique ou générale. A une échelle plus réduite, un risque organisationnel existe avec la question de la gestion du budget. Il doit être adapté aux besoins et maintenu régulièrement pour assurer un bon fonctionnement de l'ensemble des services mobilisés. De même, le facteur humain ne doit pas être négligé, l'erreur humaine reste un risque à prévoir. Ensuite, les technologies doivent être maîtrisées et une veille doit être exercée afin d'assurer que les supports physiques soient bien conservés, tout comme les formats de fichier. L'obsolescence informatique doit être prise en compte pour anticiper la perte de données. Cela passe par une maintenance du matériel mais aussi son renouvellement anticipé. Enfin, il est préférable d'avoir une bonne gestion des droits d'accès aux documents, à travers une hiérarchisation des accès. Cela permet d'avoir un meilleur contrôle de l'ensemble de la chaîne de traitement des documents à travers une hiérarchisation des responsabilités[33].

L'archivage et la conservation du patrimoine nativement numérique ont nécessité la réévaluation des outils et moyens techniques dont disposaient les bibliothèques jusqu'alors. Internet Archive s'est penché très tôt sur la question et a créé Heritrix en 2003, en collaboration avec l'IIPC. Il s'agit d'un *crawler* qui permet de collecter et d'indexer des pages web automatiquement tout en permettant une prise en main manuelle si besoin. Ce logiciel libre est majoritairement utilisé par les institutions responsables de l'archivage du Web[34]. D'autres logiciels existent pour l'archivage du Web. La British Library par exemple a eu recours

[32] Ibid. p. 20-21.

[33] Ibid. p. 21.

[34] MUSIANI Francesca, *et al.*, op. cit. p. 24.

aux logiciels PANDAS et Web HTTrack lors de sa collecte d'urgence en 2007 suite aux attentats de Londres[35].

L'archivage du patrimoine nativement numérique et par extension du Web a trouvé des solutions pour répondre aux difficultés techniques qui se présentaient. En parallèle, un autre problème de taille s'est manifesté très tôt : la question des droits d'auteur et de l'encadrement juridique des collectes.

### 1.2.3 Quels encadrements juridiques ?

Collecter et archiver le web pour sa valeur patrimoniale est certes louable mais la question des droits d'auteur s'est très vite posée. Comment respecter le droit d'auteur sur le Web alors qu'il n'est pas toujours évident d'identifier le propriétaire d'une page ou l'auteur d'un contenu. Et qu'en est-il de l'accessibilité des données collectées ? Elles dépendent de la protection des données personnelles, on ne peut les rendre accessibles à tout un chacun sans encadrement juridique. Les droits d'auteur ont ainsi représenté le premier obstacle juridique et éthique de l'archivage du Web. Si la fondation Internet Archive contourne le problème en invoquant une responsabilité civique dans l'accessibilité publique de ces données et ne les retire que lorsque l'ayant droit se manifeste[36], l'essentiel des institutions en charge d'archiver le Web ont cherché à définir un cadre légal à leurs collectes et leur accessibilité.

Les pages web sont régies par le code de la propriété intellectuelle. En temps normal, il faudrait impérativement demander à l'auteur son accord avant de collecter la moindre information. Or, ce n'est pas si simple. D'une part, il n'est pas toujours possible de retrouver l'auteur d'un contenu web. D'autre part, le web ne répond pas à une juridiction internationale unique. Chaque pays applique sa propre juridiction sur son domaine. Il faudrait alors à chaque demande pour l'étranger prendre en compte le droit local[37]. L'archivage du Web constitue déjà un projet titanesque de par l'envergure du Web, l'alourdir d'une surcouche juridique pour chaque élément archivé rendrait la tâche tout simplement impossible.

#### Une extension du dépôt légal

La plupart des pays qui ont lancé un programme d'archivage du Web ont opté pour la solution du dépôt légal. Il s'agit généralement d'une extension du dépôt légal du livre pour y

[35] SCHAFER Valérie, « Le patrimoine nativement numérique des attentats en Europe : regards croisés », in *La Gazette des archives*, Paris, 2018 n° 250, p. 116-117.

[36] MUSIANI Francesca, *et al.*, op. cit. p. 39.

[37] CHAIMBAULT Thomas, op. cit. p. 17-18.

inclure le patrimoine numérique. En France, par exemple, un dépôt légal des logiciels et des bases de données existait depuis juin 1992. C'est donc cette branche du dépôt légal qui fut modifiée pour y ajouter le Web[38]. Cette extension fait partie de la loi relative au droit d'auteur et aux droits voisins dans la société de l'information, la DADVSI, promulguée le 1er août 2006[39]. Cette loi a l'avantage de prendre en considération le droit à la propriété intellectuelle, la protection des données personnelles et la complexité caractéristique du Web, avec sa temporalité propre, son hyper-connectivité et sa situation internationale. Les institutions mandataires du dépôt légal du Web ont été autorisées à copier les sites web sans avoir à demander l'autorisation préalable des éditeurs. En parallèle, pour assurer le respect des droits de ces éditeurs et propriétaires, une restriction a été mise en place pour l'accessibilité de ces données collectées. Les archives du Web sont ainsi consultables uniquement depuis un poste informatique individuel mis à disposition au sein des institutions mandatées, c'est-à-dire au sein de la BnF ou de l'Ina et dans les pôles associés (la bibliothèque Toussaint à Angers par exemple)[40]. Si une grande partie des pays ont opté assez tôt pour la création d'un dépôt légal du Web, certains ont mis plus longtemps à régler ces questions juridiques, non sans conséquence sur leurs collectes. La British Library en est un bon exemple. Jusqu'en avril 2013, la Grande-Bretagne n'avait pas instauré de dépôt légal pour le patrimoine numérique. Pour réaliser ses collectes, la British Library était obligée de demander aux propriétaires des sites web leur autorisation pour pouvoir les collecter. Or, rien ne les obligeait à accepter. Ainsi, jusqu'en 2004, la bibliothèque n'avait pu collecter et archiver que 1 800 sites sur les 6 500 qui avaient été identifiés. La bibliothèque avait d'ailleurs alerté le gouvernement sur le sujet. Elle estimait que si le cadre légal n'était pas revu, seul 0,6 % du web britannique aurait pu être collecté en 10 ans[41].

### L'accessibilité aux archives du Web

Concernant l'accessibilité des archives du Web à travers le monde, plusieurs modèles existent. Le modèle totalement libre d'accès de la fondation Internet Archive a déjà été évoqué. Le Royaume-Uni a pris le même parti, tout comme le Portugal, la Croatie ou encore

---

[38] Article L. 131-2 du Code du patrimoine.

[39] CHAIMBAULT Thomas, op. cit. p. 12.

[40] ILLIEN Gildas, « Le dépôt légal de l'internet en pratique : les moissonneurs du Web », *BBF*, 2008, n° 6, p. 21.

[41] SCHAFER Valérie, MUSIANI Francesca, BORELLI Marguerite, op. cit., p. 137. ; ILLIEN Gildas, « Le dépôt légal de l'internet en pratique : les moissonneurs du Web », p. 21.

l'Islande. Sur le même principe d'accès restreint que la France, on peut citer l'Allemagne. L'accès n'y est possible que depuis les salles de lecture de la Bibliothèque nationale. L'Estonie est un cas intéressant d'évolution du traitement de ces archives. En 2006, une loi avait rendu accessible librement la totalité des collections numériques. Cette loi fut remplacée en 2017 par une restriction pour un meilleur respect des droits d'auteur. Ainsi, seuls les sites dont les ayants droits ont accepté la mise en ligne sont accessibles au public. Enfin, l'Espagne a sans doute la loi la plus stricte sur le respect des droits d'auteur, car elle impose le respect total du droit d'auteur pour la mise en ligne[42].

Si la question de la protection des données personnelles et de l'encadrement juridique des archives du Web semble réglée, ces lois ne sont pas immuables. Le projet de loi européenne pour formaliser le droit à l'oubli numérique avait inquiété les archivistes et bibliothécaires en 2013. En autorisant un droit à l'oubli numérique étendu, on aurait autorisé la destruction et suppression de documents physiques ou numériques au sein des archives. L'association des archivistes français s'était d'ailleurs mobilisée contre cette loi[43].

La conservation du patrimoine numérique est aujourd'hui ancrée dans les pratiques de conservation du patrimoine. Une grande partie des défis techniques et juridiques identifiés très tôt a pu être surmontée. Cependant tout comme le patrimoine physique, ce jeune patrimoine numérique en perpétuelle expansion reste très complexe à appréhender et à mesurer, d'autant plus que nous n'avons que très peu de recul sur le numérique[44]. L'encadrement progressif de sa conservation a permis la multiplication des initiatives nationales d'archivage du Web à travers le monde. Le Web ne connaissant cependant que très peu de frontières, il a fallu diversifier les approches et techniques grâce notamment à la coopération internationale.

## 2. L'archivage du Web dans le monde

L'archivage du Web s'est petit à petit imposé comme une nécessité pour la conservation du patrimoine dans une grande partie du monde. A ce jour, on ne compte pas moins de 90 projets d'archivage, menés soit par des institutions publiques nationales, soit par des institutions universitaires soit par des acteurs privés. Il n'est pas question de dresser ici une liste exhaustive de ces projets mais plutôt de s'intéresser aux initiatives publiques à visée

[42] MUSIANI Francesca, *et al.*, op. cit. p. 20.

[43] Ibid. p. 23.

[44] Ibid. p. 24.

nationale. Cela permettra de s'intéresser à la coordination internationale qui s'est développée tout autour et aux différentes techniques de collecte qui ont été déployées.

## 2.1. La multiplication des projets…

La répartition des projets d'archivage du Web n'est pas égale sur le globe. La majorité des initiatives ont vu le jour sur le sol européen, tandis que quelques projets nationaux ont été créés dans le reste du monde. Plusieurs pays bénéficient par ailleurs d'une plus grande densité d'organismes de collecte, plus ou moins spécialisés dans un domaine.

### 2.1.1. Les projets européens

L'Europe compte à elle seule 30 programmes d'archivage du Web à l'échelle nationale. Certains pays ont divisé cette mission entre plusieurs institutions, comme l'a fait la France en répartissant le dépôt légal du Web entre l'Ina pour l'audio-visuel et la BnF pour le reste. L'essentiel de ces projets a été créé entre 2000 et 2010, même si on peut aller jusqu'en 2017 pour les plus récents. La première archive du Web créée en Europe est celle de la Suède avec KulturarW[3] et il faut attendre 2000 pour voir de nouvelles bibliothèques nationales se lancer dans ce projet. La Bibliothèque nationale de République Tchèque lance alors son projet *Webarchiv* avec pour mission de collecter le Web national et de constituer des collections thématiques et qualitatives. 2003 voit la création de deux projets : d'une part en Allemagne avec la Bibliothèque centrale du Bade-Wurtemberg qui collecte et archive des sites en lien avec l'activité du Land, d'autre part au Royaume-Uni avec le projet UKGWA, *UK Government Web Archive*. Les archives du royaume s'appliquent à archiver les sites Web du gouvernement à intervalles réguliers[45].

À compter de 2004, plusieurs projets nationaux apparaissent. On assiste ainsi à la création des archives du Web croate avec *Hrvatski arhiv weba* (HAW), les archives islandaises et le projet *UK Web Archive* qui se charge d'archiver plus globalement le Web du Royaume-Uni. Le Danemark lance en 2004 ses propres archives, *Netarkivet*, par la Bibliothèque royale. L'Allemagne connaît un nouveau projet mené par le Bundestag en 2005. La même année, la Lettonie engage elle aussi sa bibliothèque nationale dans l'archivage. 2006 marque

[45] GOMES Daniel, MIRANDA João, COSTA Miguel, « A survey on Web archiving initiatives », TPDL 2011. Proceedings of the 15th international conference on Theory and practice of digital libraries: research and advanced technology for digital libraries, Berlin/Heidelberg/New York, Springer, 2011, p. 408-420. ; Les auteurs ont créé en lien avec leur article une page Wikipédia regroupant l'ensemble des initiatives d'archivage du Web à travers le monde en 2011, notamment par les institutions elles-mêmes : https://en.wikipedia.org/wiki/List_of_Web_archiving_initiatives

l'instauration d'un cadre légal pour le dépôt légal numérique et par extension pour l'archivage du Web en France. En pratique, le décret d'application de la loi DADVSI n'arrivera qu'en 2011, mais la BnF continue ses expérimentations débutées en 1999 en attendant. L'Ina débutera l'archivage effectif du Web audio-visuel en 2009. 2007 voit le lancement de quatre projets. L'Autriche commence à archiver les périodiques et publications en lien avec la littérature avec son projet DILIMAG, *Digital Literature Magazines*. Elle a étendu son périmètre de collecte un an plus tard à l'ensemble du domaine autrichien. Le Portugal lance *arquivo.pt* qui collecte le domaine portugais mais aussi celui du Cap-Vert (.cv) et de l'Angola (.ao) au titre de l'héritage colonial et de l'absence pour le moment d'un archivage national. La Slovénie et l'Ukraine débutent également l'archivage de leur Web national cette même année. La Finlande, les Pays-Bas et la Suisse se lancent dans ce projet en 2008. L'année d'après c'est au tour de la Serbie et de l'Espagne. L'Espagne avait déjà en pratique deux projets en activité : le web catalan depuis 2005 avec le projet PADICAT et le Web basque depuis 2008 avec ONDARENET. L'Estonie, la Grèce et la Russie lancent leurs propres programmes en 2010, suivis l'année suivante par l'Irlande. L'Union Européenne crée l'*EU web archive* en 2013 pour archiver les sites des institutions européennes. En 2015, le Luxembourg et la Slovaquie archivent à leur tour le Web. Enfin, 2017 voit les initiatives les plus récentes avec la Hongrie et la Belgique[46]. Ces deux projets sont encore des pilotes qui s'inspirent de l'expérience des autres institutions à l'international. Le projet belge PROMISE est mené par la Bibliothèque royale et les Archives nationales, en collaboration avec plusieurs universités[47].

### 2.1.2. Les projets extra-européens

En dehors des frontières de l'Europe, on compte 12 pays bénéficiant d'au moins un programme d'archivage du Web. Les États-Unis ont été pionniers dans le domaine avec Internet Archive. Ils restent le pays avec la plus grande densité de projets, on n'en compte pas moins de 23. La plupart sont menés par des entreprises privées. Des universités se sont cependant impliquées dans le domaine, tout comme des institutions publiques comme la Bibliothèque du Congrès qui collecte et archive le Web national depuis 2000. L'Australie fait également partie des pionniers avec PANDORA.

Dès 1999, la Nouvelle-Zélande a commencé à archiver son domaine national. Une partie de l'Asie s'intéresse aussi à la conservation du Web. Ainsi la Corée du Sud lance en 2001 OASIS, *Online Archiving & Searching Internet Sources*, tandis que le Japon inaugure en 2002 le WARP, *Web Archiving Project*, piloté par la Bibliothèque nationale de la Diète. La même

[46] https://en.wikipedia.org/wiki/List_of_Web_archiving_initiatives

[47] MUSIANI Francesca, et al., op. cit., p. 19.

année, l'Égypte inaugure la *Bibliotheca Alexandrina* et le site miroir d'Internet Archive. La Chine débute l'archivage de son domaine national en 2003 avec le projet WICP, *Web Information Collection and Preservation*. En 2005, la Bibliothèque et Archives Canada lance un premier programme d'archivage que la Bibliothèque et Archives nationales du Québec est venue compléter en 2012. Toujours en 2006, Singapour confie cette mission d'archivage à sa bibliothèque nationale. Taiwan fait de même l'année suivante. Enfin, Israël débute son propre programme en 2011[48].

L'archivage du web s'étend de plus en plus à travers le monde, avec un réel essor depuis 2002. Toutefois, on constate que dans ce domaine également, une fracture Nord/Sud est bien présente.

### 2.1.3. Une forte disparité Nord/Sud

La fracture numérique est un problème bien connu. Entre le Nord, principalement les pays occidentaux, qui bénéficie d'une couverture numérique et Internet maximale et le Sud qui reste encore largement délaissé[49]. Les capitales et villes principales de ces pays profitent d'une couverture relativement bien développée, tandis que les territoires plus ruraux restent pour la plupart sans accès à Internet. Au-delà de la question de la mise à disposition du réseau Internet, il faut également avoir accès à du matériel informatique, ce qui représente un coût supplémentaire.

L'archivage du Web et par extension la conservation du patrimoine numérique demeure à ce jour une préoccupation essentiellement pour les pays du Nord[50]. On compte deux initiatives de pays en développement. D'une part en Égypte, qui se trouve cependant parmi les pays les plus développés de cette catégorie, et d'autre part le Chili qui a débuté un programme d'archivage et intégré l'IIPC. On constate que certains pays comptent de nombreux programmes simultanés qui ont des objectifs variés.

Force est de constater que les pays en développement s'intéressent encore très peu aux problématiques de l'archivage du Web. L'IIPC l'a mis en valeur suite à l'échec de sa campagne de communication et de sensibilisation auprès des pays en développement, entre 2007 et 2009. La conservation du patrimoine numérique ne représente nullement une priorité lorsque le fait même d'avoir des bibliothèques et livres en nombre suffisant est un problème, ou que l'accès à l'électricité n'est pas garanti à tous. En ce sens, des initiatives comme celle d'Internet Archive, qui collecte le web mondial, permet de minimiser la perte du Web de ces pays, tout

---

[48] https://en.wikipedia.org/wiki/List_of_Web_archiving_initiatives

[49] SCHAFER Valérie, MUSIANI Francesca, BORELLI Marguerite, op. cit., p. 134.

[50] Annexe 1 : carte des pays ayant au moins un programme d'archivage du Web.

en leur permettant un accès direct et gratuit grâce à l'*open source*, en attendant que des programmes d'archivage nationaux soient créés[51].

Pour pouvoir partager leurs expériences et assurer une coordination internationale des programmes d'archivage du Web, des collaborations internationales ont vu le jour. Certaines ont pris fin, d'autres, comme l'IIPC, continuent leurs activités.

## 2.2. … Et la collaboration internationale

Avec la multiplication des programmes d'archivage du Web dans plusieurs pays, l'idée de se réunir pour partager les différentes expériences et évoluer ensemble a très rapidement germé. Plusieurs organisations internationales ont été créées depuis 2003, la plus importante reste néanmoins l'IIPC.

### 2.2.1. IIIPC : International Internet Preservation Consortium

Dès le début de l'archivage du Web, la question du périmètre de collecte de chaque programme s'est posée. Si Internet Archive contourne le problème en optant pour l'approche intégrale, les autres institutions s'intéressant au sujet ont dû définir leur périmètre d'action. Or, Internet est international, dans sa gouvernance comme dans ses contenus. On ne pouvait pas définir le Web national en se basant uniquement sur des critères valables pour les archives ou l'édition jusqu'alors, comme la langue, la nationalité du créateur de contenu ou bien les frontières physiques. Internet recoupe en permanence ces critères. Par exemple, comment définir quelle partie de YouTube archiver pour un pays. Ce questionnement a mis en lumière le besoin d'une interopérabilité entre les collections de chaque initiative. En combinant les collectes faites dans des périmètres restreints par chaque pays, on s'assurerait une meilleure couverture globale du Web, une plus grande finesse dans le degré de collecte. Chacun pourrait alors compléter ses lacunes documentaires avec les collections des autres[52]. Plusieurs pistes ont été explorées. Par exemple, Brewster Kahle a proposé en 2003 de créer une archive mondiale du Web unique qui serait enrichie par les bibliothèques nationales qui y archiveraient leur Web national. Les questions juridiques posaient cependant problème au regard de la diversité des lois sur le droit d'auteur et la protection des données personnelles dans chaque pays. En juillet de la même année, une solution, sous forme de compromis, a été trouvée.

---

[51] ILLIEN Gildas, « Une histoire politique de l'archivage du web. Le consortium international pour la préservation de l'Internet », BFF, 2011, n° 2, p. 65.

[52] Ibid. p. 62. ; MUSSOU Claude, « Et le Web devint archive : enjeux et défis », Le Temps des médias, 2012, vol. 19, n° 2, p. 260-261.

Chaque institution continuera d'archiver selon ses propres approches son périmètre du Web tout en travaillant en collaboration avec toutes les autres pour assurer l'interopérabilité des collections, éviter les redondances et la dispersion des efforts[53]. L'IIPC a été fondée par douze membres, soit onze bibliothèques nationales et Internet Archive. On retrouve donc la Bibliothèque nationale d'Australie, la Bibliothèque et archives nationales du Canada, la Bibliothèque Royale du Danemark, la Bibliothèque nationale de Finlande, La Bibliothèque nationale de France, la Bibliothèque nationale d'Islande, la Bibliothèque nationale d'Italie, la Bibliothèque nationale de Norvège, la Bibliothèque Royale de Suède, la British Library et enfin la Bibliothèque du Congrès[54]. Elle compte aujourd'hui cinquante-six membres[55].

### Des objectifs ambitieux

Lors de la création officielle de l'IIPC à Paris en 2003, les membres fondateurs ont défini trois missions primordiales qui cernent les problématiques autour de l'archivage du Web :

*« - Travailler en collaboration, dans le cadre législatif de leurs pays respectifs, pour identifier, développer et faciliter la mise en œuvre de solutions permettant de sélectionner, de collecter et de préserver les contenus de l'internet et d'en assurer l'accessibilité.*

*- Faciliter la couverture internationale des collections d'archives de contenus de l'internet, en conformité avec leurs cadres législatifs nationaux et en accord avec leurs politiques respectives de développement des collections nationales.*

*- Plaider vigoureusement au niveau international en faveur d'initiatives et de lois encourageant la collecte, la préservation et l'accès aux contenus de l'internet[56]. »*

Pour s'assurer que ces missions soient suivies et qu'elles restent au cœur des réflexions au fil des évolutions du consortium, quatre engagements ont également été ratifiés :

*« - Offrir un forum pour le partage des connaissances sur l'archivage des contenus de l'internet ;*

*- Développer et promouvoir des normes pour la collecte, la préservation et l'accès à long terme aux contenus de l'internet ;*

---

[53] ILLIEN Gildas, « Une histoire politique de l'archivage du web. Le consortium international pour la préservation de l'Internet », p. 62.

[54] CHAIMBAULT Thomas, op. cit., p. 37.

[55] http://netpreserve.org/about-us/members/

[56] Cf. ILLIEN Gildas, « Une histoire politique de l'archivage du web. Le consortium international pour la préservation de l'Internet », p. 63.

*- Favoriser le développement de logiciels et d'outils appropriés et interopérables, de préférence sous licence libre (open source) ;*

*- Améliorer la sensibilisation aux questions liées à la préservation des contenus de l'internet et aux initiatives associées, notamment par le biais de conférences, d'ateliers, de formations, de publications*[57]*. »*

L'idée derrière ce consortium est bien de permettre la meilleure collaboration possible entre les programmes d'archivage du Web. Il s'agit aussi de leur fournir tous les outils dont ils peuvent avoir besoin et de réfléchir conjointement aux meilleures solutions d'archivage et de stockage. Il a été décidé de traiter ces engagements dans l'ordre de leur priorité. L'accent a donc été mis entre 2003 et 2006 sur la création d'un *crawler* répondant aux besoins et exigences des institutions. Ceux disponibles alors n'étaient pas assez performants par rapport au besoin de collecter et traiter automatiquement un important volume de données et de sites. Un appel d'offre a d'abord été lancé par la British Library, la Bibliothèque du Congrès et la BnF afin d'en assurer une meilleure visibilité. Malgré cela, les trois appels sont restés sans réponse. Cela n'a pas pour autant découragé les membres de l'IIPC qui ont donc décidé de créer eux-mêmes les outils dont ils avaient besoin. Dans l'optique de garantir l'accès à tous et l'adaptabilité aux besoins de chacun, le choix de développement s'est porté sur l'*open source*. Ils ne sont cependant pas partis de rien, les premiers tests réalisés par Internet Archive jusqu'alors ont été mis à profit[58]. De ce projet est né Heritrix, un des *crawler* les plus utilisés encore aujourd'hui, notamment par la BnF et Internet Archive, et est compatible avec la *Wayback Machine*[59].

En parallèle du développement d'Heritrix, des groupes de réflexion ont été lancés sur la question des formats et de leur compatibilité sur divers supports. Ces questionnements ont été placés au cœur des recherches de l'IIPC entre 2007 et 2009. Le Web est composé d'une mosaïque de formats qui rendent plus difficile l'interopérabilité entre les collections. Chaque institution peut archiver et stocker ses données et fichiers sous le format qu'elle souhaite. L'IIPC a donc cherché un moyen de normaliser le format de stockage des données. Le format ARC, créé par Brewster Kahle dans les années 1990, a servi de base, et a abouti au format WARC, pour *Web ARChive*. Ce nouveau format permet de traiter et gérer de plus gros volumes de données que son prédécesseur. Grande innovation, il permet également de sauvegarder directement les métadonnées associées aux pages archivées, ce qui favorise le traitement, l'archivage et la consultation de ces pages, ouvrant la voie plus largement à la consultation des

[57] Cf. Ibid. p. 63.

[58] Ibid. p.62. ; GEBEIL Sophie, « Pourquoi archiver le Web ? Les missions de l'IIPC », *Carnet de recherche Internet, histoire et mémoires,* 2014. https://madi.hypotheses.org/243

[59] GEBEIL Sophie, op. cit.

archives du Web. Le format WARC a été normalisé par l'ISO en 2009. En 2010, l'IIPC a également adopté le modèle de stockage OAIS (*Open Archive Information System*) qui permet de retrouver des données et des fichiers au sein des archives[60]. L'IIPC continue de travailler à perfectionner ses outils et à répondre aux besoins techniques. Au fur et à mesure, une boîte à outil a été créée et proposée à toutes les institutions qui le souhaitent, grâce au choix de l'*open source*.

### L'organisation

Comment peut fonctionner une organisation internationale souhaitant créer tous ces outils et normes ? L'IIPC fonctionne sur la base du volontariat et bénévolat. Il n'a aucun salarié permanent, n'a pas de siège et ses membres ne se rencontrent en présentiel qu'une à deux fois par an. Chaque nouvelle adhésion se fait par cooptation et nécessite de verser une cotisation annuelle définie selon le budget de l'institution. Cela peut aller de 2 000€ à 8 000€. Au début de l'IIPC, c'est la BnF qui a piloté les projets. Puis progressivement différents pôles ont été définis pour répartir la gouvernance. Ainsi, la Bibliothèque du Congrès a pris en charge la communication depuis 2007 tandis qu'Internet Archive gère le pilotage technique depuis 2010. La présidence est quant à elle soumise au vote annuellement. Pour pallier les difficultés dues à l'éloignement géographique, des conférences téléphoniques sont organisées régulièrement pour faciliter les échanges. Cela permet aussi au comité de pilotage de voter plus régulièrement les décisions importantes. Ensuite, plusieurs groupes de travail ont été créés pour diversifier les projets tout en évitant l'éparpillement. Chaque groupe est dirigé par un binôme issu de deux institutions différentes. Trois grands axes ont été définis pour ces groupes : la collecte, la préservation et l'accès, ce qui permet d'encadrer les missions et engagements de l'IIPC[61].

L'IIPC continue de travailler à améliorer ses outils et lance chaque année de nouveaux groupes de travail et projets. En 2018, l'accent a été mis sur l'amélioration et la favorisation des échanges entre les membres. Ces projets visent aussi bien les institutions en tant que telles mais aussi les compétences personnelles et professionnelles des personnes impliquées dans ces projets[62]. Enfin, malgré une ouverture internationale et des appels à contribution

---

[60] Ibid. ; MUSIANI Francesca, et al., op. cit., p. 16-61.

[61] ILLIEN Gildas, « Une histoire politique de l'archivage du web. Le consortium international pour la préservation de l'Internet », p. 63.

[62] http://netpreserve.org/projects/

auprès de pays en développement, on constate que la fracture Nord/Sud est très présente parmi les membres du consortium[63].

### 2.2.2. NEDLIB : Networked European Deposit Library

L'IIPC n'est pas le seul programme de coopération internationale autour de l'archivage du Web, certains, comme le NDIIPP (*National Digital Information Infrastructure and Preservation Program*) aux États-Unis s'organisent au niveau national. Il coordonne les différents programmes de préservation du numérique à travers tout le pays. Son fonctionnement a par ailleurs été calqué sur celui de l'IIPC[64]. Le Royaume-Uni a lui aussi développé un programme de coordination de ses archives du Web, qui regroupe la British Library, les archives nationales, le JISC (Joint Information Systems Committee), la Bibliothèque nationale d'Écosse, la Bibliothèque nationale du Pays de Galles et Wellcome Trust, en lien avec le monde médical et de la santé. Chacun prend en charge quelques domaines dont il a une meilleure connaissance.

NEDLIB était un projet lancé par la Commission européenne dans les premières années de l'archivage du Web. Il a débuté en 1998 avec pour mission de développer les infrastructures nécessaires à la création d'un dépôt collaboratif européen et pour aider les institutions européennes à disposer d'outils de collecte. Au total, onze bibliothèques et trois éditeurs ont participé au projet. Ainsi, l'Allemagne, la Finlande, la France, l'Italie, la Norvège, les Pays-Bas, le Portugal et la Suisse se sont alliés à Kluwer Academic, Elsevier Science BV et Springer-Verlag. De cette collaboration est né le *crawler* NEDLIB. Il a cependant été par la suite remplacé par Heritrix, plus performant et mieux adapté. Le projet NEDLIB a pris fin en 2001, après avoir atteint ses objectifs[65].

## 2.3. Des techniques de collecte diversifiées

Comme cela a déjà été évoqué, il existe plusieurs approches de collectes. Chaque institution adopte et adapte un modèle pour répondre à ses besoins, à sa politique documentaire ainsi qu'au cadre juridique qui encadre sa mission d'archivage. On compte trois

---

[63] Annexe 2 : carte de la répartition des membres de l'IIPC.

[64] ILLIEN Gildas, « Une histoire politique de l'archivage du web. Le consortium international pour la préservation de l'Internet », p. 65.

[65] CHAIMBAULT Thomas, op. cit., p. 38. ; https://cordis.europa.eu/project/rcn/43331/factsheet/fr

approches de collecte majeures et quelques autres plus rares et souvent complétées avec une des trois premières.

### 2.3.1. L'approche intégrale

L'approche intégrale représente un idéal de préservation patrimoniale pour les archives du Web. Avec ce type de collecte, on moissonne l'intégralité du Web sans le moindre regard sur l'origine du document ou du site, ni son intégrité ou encore sa qualité. Cela implique évidemment de ne pas prendre en considération la question du droit d'auteur ou de la protection des données personnelles au titre d'une vision patrimoniale du Web[66].

Seul Internet Archive pratique l'approche intégrale. Avec l'application du *opt-out*, il peut collecter l'ensemble du Web mondial. Si un ayant-droit se manifeste, les éléments collectés sont simplement retirés de l'accès public en ligne mais toujours conservés dans leurs archives.

### 2.3.2. L'exhaustivité automatisée

L'approche exhaustive permet de collecter en appliquant quelques critères de sélection. La plupart du temps, il s'agit d'une restriction des noms de domaines pour ne collecter que le Web national. En restreignant ainsi le périmètre de collecte, on peut s'assurer la collecte d'un volume assez important d'URL. Du fait de ce volume, l'automatisation est majoritairement privilégiée, plutôt qu'une collecte manuelle ou semi-automatisée. Cela permet de limiter le coût de la procédure ainsi que sa durée, le *crawler* étant bien plus rapide.

Le programme KulturarW$^3$ suédois a choisi cette approche. Il peut ainsi moissonner l'intégralité des noms de domaine en .se et ceux édités sur le territoire et ainsi couvrir une grande partie du Web national. La Finlande applique elle aussi cette approche mais va plus loin dans la restriction du périmètre. Elle ne collecte que le domaine national en .fi et ignore les .com par exemple. Un faible échantillon des autres domaines est malgré tout collecté à travers les liens capturés lors du moissonnage, comme les vidéos, les images, les publications liées, etc[67].

---

[66] CHAIMBAULT Thomas, op. cit., p. 26. ; GHARSALLAH Mehdi, op. cit., p. 6.

[67] CHAIMBAULT Thomas, op. cit., p. 26. ; SCHAFER Valérie, MUSIANI Francesca, BORELLI Marguerite, op. cit., p. 138.

### 2.3.3. L'échantillonnage semi-automatisé

L'approche par échantillonnage semi-automatisé permet une plus grande finesse dans le résultat final de la collecte. Une pré-sélection de sites à collecter, jugés intéressants pour leur contenu, leur qualité ou encore leur représentativité, est établie par des bibliothécaires. Cette liste est ensuite collectée automatiquement par le *crawler*, à intervalles fréquents.

Le projet PANDORA de la Bibliothèque nationale d'Australie est un bon exemple de la mise en pratique de cette approche, avec un protocole bien défini. Le contenu et la structure d'un site sont évalués en amont et si le site est sélectionné, une demande de permission de collecte est envoyée à l'éditeur de la publication. Une fois l'accord donné, la publication est cataloguée dans la base de données de la Bibliothèque nationale d'Australie, notamment pour créer un lien hypertexte afin d'en assurer l'accès. Une requête est ensuite envoyée pour lancer la collecte par le *crawler*. Une fois le site collecté, on vérifie manuellement si l'archive est conforme au site en ligne pour s'assurer que toutes les pages et informations ont été capturées. Un rapport de vérification est transmis, soit pour confirmer que l'archive est une copie conforme, soit pour signaler les erreurs à corriger. Une fois l'archive validée, on établit une page d'entrée dans le catalogue pour la page archivée et on lui attribue une PURL, *Persistent Uniform Ressource Locator*, qui assure une meilleure stabilité qu'une URL car elle suit les changements d'adresse URL éventuels, ce qui évite les erreurs 404, entre autres. Enfin, pour assurer la pérennité de ces archives, on fait une vérification périodique pour vérifier l'état de l'archive et la comparer avec le site encore en ligne et compléter éventuellement s'il y a eu des modifications[68].

### 2.3.4. Autres approches

Il existe d'autres approches que celles précédemment évoquées. Elles sont plus minoritaires car bien souvent elles ne collectent qu'un périmètre très restreint. On parle alors d'approche thématique et d'approche disciplinaire.

L'approche thématique s'inscrit généralement dans une démarche exceptionnelle. Elle concerne une collecte en lien avec un événement particulier par exemple. Il s'agit alors de faire une sélection de sites s'y rapportant pour constituer une archive plus précise et détaillée autour de cet événement qui aura sans doute été survolé, voir ignoré lors de la collecte exhaustive. La BnF réalise des collectes thématiques à l'occasion des élections nationales par exemple.

L'approche disciplinaire, quant à elle, permet de se concentrer sur une discipline en particulier. La plupart du temps, elle est employée par les institutions universitaires ou de

[68] GHARSALLAH Mehdi, op. cit. p 6.

recherche qui réalisent des archives autour de leur domaine. Par exemple, l'université de Heidelberg en Allemagne collecte les sites en rapport avec les études chinoises avec son projet DACHS, *Digital Archives for Chines Study*. Le projet néerlandais ARCHIPOL s'intéresse de son côté à la politique[69].

En pratique, la plupart des programmes d'archivage du Web à l'échelle nationale combine plusieurs approches afin de couvrir le maximum de surface du Web national. La BnF par exemple pratique l'exhaustivité pour sa collecte large annuelle et réalise des collectes sélectives régulières pour approfondir ses collections, tout comme elle réalise des collectes thématiques. La mise en relation des différents programmes à travers le monde a permis de partager toutes ces approches, leurs avantages et leurs inconvénients, pour que chacun puisse affiner son protocole au sein de son propre programme. Cette coopération mondiale a permis à la France de développer progressivement ses procédures de collecte pour être opérationnelle dès la promulgation de la loi DADVSI.

# 3. L'archivage du Web en France, une répartition entre deux institutions

Depuis François Ier et l'ordonnance de Montpellier du 28 décembre 1537, « *l'obligation pour tout éditeur, imprimeur, producteur, importateur de déposer chaque document qu'il édite, imprime, produit ou importe en France à la BnF ou auprès de l'organisme habilité à recevoir le dépôt en fonction de la nature du document*[70] » a été instauré. Au fil du temps, le dépôt légal s'est étendu pour intégrer de nouveaux supports comme les estampes, partitions, cartes et plans, photographies, affiches ou l'audiovisuel. Dès la fin des années 1990, le Conseil d'État s'est penché sur le contenu créé et mis à disposition sur le Web et a demandé un rapport sur l'Internet et les réseaux numériques. On y évoque dès le premier article que « *l'Internet stimule fortement la créativité littéraire et artistique comme en témoigne la croissance très rapide des sites et des pages personnelles. Il en résulte une multiplication des œuvres mises à la disposition du public et donc en principe assujetties au dépôt légal. Si l'on souhaite que cette obligation reste respectée, il faut en faciliter les modalités d'exécution. Il serait en particulier indispensable que le dépôt des œuvres puisse désormais se faire en ligne*[71] ». Partant de cette

[69] CHAIMBAULT Thomas, op. cit., p. 27.

[70] Cf. MUSIANI Francesca, *et al.*, op. cit., p. 17.

[71] Cf. FRANCE (CONSEIL D'ETAT), *Internet et les réseaux numériques*, Paris, La Documentation française, 1998.

base, la réflexion autour de l'extension du dépôt légal pour y inclure le Web débute. Elle aboutira à la promulgation de la loi DAVDSI le 1er août 2006. Une approche très ouverte a été favorisée pour assurer la prise en compte de la globalité du Web présent et de ses évolutions à venir. Il est donc stipulé que « sont également soumis au dépôt légal les signes, signaux, écrits, images, sons ou messages de toute nature faisant l'objet d'une communication au public par voie électronique[72] ». Ainsi, sont exclus du dépôt légal uniquement les correspondances et espaces privés hébergés sur le Web[73], au titre de la protection des données personnelles. Le décret d'application du 19 décembre 2011 a marqué le début officiel de l'archivage du Web en France. Cette mission a été répartie entre l'Ina et la BnF.

## 3.1. L'Institut national de l'audiovisuel

Depuis 1992, l'Ina est en charge du dépôt légal de l'audiovisuel français et trois ans plus tard les diffuseurs hertziens y ont déposé leurs émissions, puis cela a été le tour des chaînes câblées et satellites puis les radios privées. La captation directe des flux d'émissions instaurée par la suite a permis une plus grande efficacité dans la collecte de l'Ina.

### 3.1.1. Le périmètre du dépôt légal du Web

Avec la loi DADVSI, l'Ina a vu son périmètre de collecte élargi[74]. Elle collecte ainsi « les sites émanant des services des médias audiovisuels, Web Tv et Web radios, les sites principalement consacrés aux programmes radio et télé, les sites des organismes de l'environnement professionnel et institutionnel du secteur de la communication audiovisuelle[75] ». La loi DADVSI est entrée en application fin 2011 mais en pratique, l'Ina avait déjà commencé à collecter et archiver le Web pour expérimenter et être pleinement opérationnel lors de la mise en application légale. Ainsi, dès 2001, l'Ina avait commencé à collecter les sites de médias audiovisuels et de sites en lien avec l'audiovisuel, venant compléter la première sélection[76].

---

[72] Cf. article L. 131-2 du Code du patrimoine.

[73] GAME, Valérie, OURY, Clément, « Le dépôt légal de l'Internet à la BnF : adapter une mission patrimoniale à l'économie de l'immatériel », in Institut national du Patrimoine, *Le patrimoine culturel au risque de l'immatériel : enjeux juridiques, culturels, économiques : actes du colloque [colloque, Paris, INP, 3-4 avril 2008]*, Paris, L'Harmattan, 2010, p. 60-63.

[74] HOOG Emmanuel, *L'Ina*, Paris, PUF, 2006, p. 96.

[75] https://institut.ina.fr/institut/statut-missions/depot-legal-radio-tele-et-web

[76] MUSSOU Claude, op. cit., p. 261.

En plus de ces sites, les réseaux socionumériques ont pris une place prépondérante dans la production de contenu, y compris pour l'audiovisuel. L'Ina collecte donc depuis 2016 12 600 comptes Twitter et 7 800 comptes de plateformes vidéo comme des chaînes YouTube. Pour ces dernières, la collecte a commencé à être mise en place en 2008. Pour compléter les comptes Twitter, 400 hashtags sont également suivis, tous en lien avec des acteurs clés de l'audiovisuel français. Twitter étant un réseau très dynamique, une veille a été instaurée pour permettre de ne pas rater les principales tendances de mots-clés, les *trends*, même lorsque le service est fermé, la nuit et les weekends. Une collecte automatique est programmée pour récupérer les principales tendances[77].

Pour collecter ces tweets, l'Ina n'utilise pas Heritrix, qui n'est pas le *crawler* le mieux adapté à la capture du contenu vidéo ou image, mais l'API, interface de programmation, mis à disposition gratuitement par Twitter. Cet API se divise en deux outils : l'API *Search* et l'API *Streamin*g. L'API *Search* permet de remonter jusqu'à un contenu particulier, dans la limite des sept derniers jours, tandis que l'API *Streaming* permet de capter un flux au fur et à mesure. Toutefois, cet outil gratuit a ses limites. Il ne permet de collecter que 1 % des tweets émis à l'échelle mondiale à un instant T. Cependant, le but de ces collectes n'est pas la recherche de l'exhaustivité mais plutôt de la représentativité du Web français. Ces 1 % peuvent suffire si le choix des hastags est élaboré en amont[78].

### 3.1.2. L'accessibilité des archives

Pour respecter le droit d'auteur, l'Ina restreint l'accès à l'ensemble des programmes dont elle ne bénéficie pas des droits d'exploitations. Ces documents sont archivés au même titre que les autres mais ils ne sont accessibles que sur demande et présentation d'un travail de recherche qui justifie de la nécessité de cette consultation. Des outils de navigation et d'exploitation des données ont par ailleurs été développés pour aider les chercheurs dans leurs travaux. Pour répondre au mieux à ces besoins, l'Ina a lancé en 2009 des ateliers sur les usages de la recherche du Web et de son archive. Pour compléter ces travaux, une enquête à été menée entre le 5 et le 30 mai 2011 au sein de réseaux de chercheurs. L'enquête « Un Web archivé pour quoi faire ? » a mis en lumière la forte présence des sciences humaines dans ces recherches. Cependant, les chercheurs ont manifesté le besoin d'outils proches de ceux utilisés pour les sciences et technologies de l'information et de la communication[79].

---

[77] https://institut.ina.fr/institut/statut-missions/depot-legal-radio-tele-et-web ; MUSIANI Francesca, *et al.*, op. cit., p. 43-44.

[78] Ibid. p. 44-45.

[79] MUSSOU Claude, op. cit., p. 262.

Enfin, pour permettre une bonne accessibilité de ses archives auprès des chercheurs mais aussi du grand public, l'Ina a déployé tout un réseau de centres de consultation sur le territoire français. Deux niveaux de possibilité de consultation ont été établis. Le premier, destiné aux chercheurs, est appelé consultation experte. Elle permet d'accéder à l'intégralité des fonds de l'Ina tout en bénéficiant de l'accompagnement et de l'expertise des équipes. Des outils d'aide à l'analyse sont également disponibles. On compte sept centres de consultation de ce type, situés à Lille, Lyon, Marseille, Rennes, Strasbourg, Toulouse et Paris BnF. Le second niveau, la consultation autonome, vise plutôt le grand public. Elle donne accès aux notices descriptives de l'ensemble des fonds de l'Ina et permet le visionnage et l'écoute des fonds TV, radio et Web média. La navigation se fait sur une plateforme intuitive, qui propose une recherche simple ou multicritères et une recherche dédiée au Web média, à travers le nom du site, l'URL, la recherche plein texte et la date. Cet accès se fait depuis les bibliothèques municipales à vocation régionale. Ainsi, on compte 43 bibliothèques réparties sur tout le territoire : Aix-en-Provence, Amiens, Angers, Avignon, Besançon, Bordeaux, Brest, Bourges, Caen, Clermont-Ferrand, deux à Dijon, Faa'a Tahiti en Polynésie française, Fort-de-France en Martinique, Gourbeyre en Guadeloupe, deux à Grenoble, La Rochelle, Le Mans, Limoges, Lyon, Metz, Montpellier, Nancy, Nantes, Nice, Nîmes, Orléans, Pau, Perpignan, Pessac, Poitiers, Porto-Vecchio, Reims, Rennes, Rouen, Saint Étienne, Saint-Martin-d'Hères, Strasbourg, Toulouse, Tours et deux à Villeneuve d'Ascq[80].

## 3.2. La Bibliothèque nationale de France

Suite à la loi DADVSI, la BnF a reçu la mission de collecter et d'archiver tout ce qui n'était pas du ressort de l'Ina. Le Web français reste cependant très vaste et il a fallu définir un périmètre de collecte pour mener à bien cette charge tout en expérimentant pour établir un protocole de collecte adapté.

### 3.2.1. Le périmètre du dépôt légal du Web

Certains pays ont choisi d'arrêter leur périmètre sur le critère de la langue ou celui du nom de domaine national. La BnF n'a pas souhaité se limiter à ces critères. Elle a privilégié la caractérisation du Web français par son lien avec le territoire français. Ainsi, la BnF collecte les sites dont l'éditeur est basé en France, même s'il n'est pas hébergé en France. Par conséquent, le critère de langue ne pourrait suffire puisqu'une partie de ce Web collecté est en langue étrangère. La Bnf collecte donc les .fr et les domaines des territoires d'outre-mer comme les

[80] https://institut.ina.fr/institut/statut-missions/depot-legal-radio-tele-et-web

.re (La Réunion) ou .mq (Martinique). Ces sites sont aisément identifiables puisqu'ils tiennent tous leur nom de domaine de l'AFNIC, l'Association française pour le nommage Internet en coopération, qui attribue et gère les domaines français. Chaque année, l'AFNIC procure à la BnF la liste complète des sites avec un nom de domaine français pour sa collecte. Cependant, la BnF ne se limite pas à cette liste. Elle moissonne également une partie des domaines en .com, .net ou encore .org basés en France. L'identification de ces sites pose plus de problèmes. Plusieurs bureaux d'enregistrement acceptent de fournir leurs listes à la BnF, mais pas tous. La loi ne permet pas non plus de les y contraindre[81].

Le dépôt légal du Web géré par la BnF, tout comme pour l'Ina, ne recherche pas l'exhaustivité. Comme la loi l'indique, il s'agit d'un échantillonnage représentatif de l'état et du contenu du Web français au moment de la collecte. En cela, c'est une grande différence avec le reste du dépôt légal des livres, périodiques, documents audiovisuels sur support, cartes et plans, etc[82]. À ce jour, la BnF estime couvrir environ 60 % du Web français. Cependant, il faut prendre avec précaution ce chiffre car il est très difficile d'évaluer avec exactitude l'ampleur du Web mondial ou national. Les estimations que l'on trouve sont par ailleurs difficilement vérifiables puisque remonter à la source est quasiment impossible[83].

### 3.2.2. Historique de l'évolution de l'archivage du Web

Bien que le décret d'application de la loi DADVSI ne soit arrivé que fin 2011, la BnF a débuté l'archivage du Web, à titre expérimental, dès le début des années 2000. Membre fondateur de l'IIPC, elle a suivi de près les évolutions de l'archivage.

#### De l'initiative à l'expérimentation : 1999-2004

Entre 1999 et 2004, la BnF s'est intéressée aux premières initiatives d'archivage du Web à travers le monde, à commencer par Internet Archive. Sous l'impulsion de Julien Masanès et Jean-Noël Jeanneney, les premières collectes ont été réalisées. Les élections présidentielles de 2002 ont servi de premier test de sélection de sites, puis l'opération a été reproduite pour les élections régionales et européennes de 2004. Ces premières collectes n'ont pas été réalisées en interne. La BnF ne disposait pas encore des moyens techniques et humains pour mener à bien ces collectes. Elle a donc mandaté Internet Archive pour les

[81] Annexe 3 : entretien à la BnF. ; GAME, Valérie, OURY, Clément, op. cit., p. 64.

[82] Annexe 3 : entretien à la BnF.

[83] Ibid.

exécuter. Au total, 3 500 sites ont été collectés lors de ces deux collectes expérimentales. La BnF s'implique également activement au sein de l'IIPC et en assure la présidence en 2003 suite à sa création.

### Stabilité acquise et encadrement juridique : 2004-2007

La période entre 2004 et 2007 est marquée par le vote de la loi DADVSI en août 2006. En préparation depuis un peu plus de cinq ans, cette loi marque le début de l'encadrement et de la reconnaissance du travail de la BnF. D'expérimentation, l'archivage du Web devient une véritable mission au sein de la BnF. D'autres collectes ont été réalisées pendant ces années, toujours avec l'aide extérieure d'Internet Archive.

En parallèle, la BnF a déployé en 2005 tout un réseau de correspondants du dépôt légal du Web. Elle enseigne à 35 personnes comment archiver les sites, évaluer les collectes. Cela a été l'occasion d'élaborer une ébauche de politique de sélection pour la réalisation de collectes ciblées, et la mobilisation des pôles associés pour ces collectes commence à être évoquée[84].

### Le premier cycle complet d'archivage : 2007-2012

Cette période a marqué le début de l'archivage dans l'enceinte même de la BnF. Pour la première fois, une collecte large complète a été réalisée sur place, par les équipes de la BnF. Jusqu'en 2008, la collecte annuelle était réalisée par Internet Archive. Pendant cette phase, la BnF a lancé des collectes courantes en interne avec des fréquences plus ou moins régulières, sur un ensemble total de 5 000 sites. Des collectes projets ont aussi été réalisées en interne, avec d'une part des collectes pour des élections et d'autre part autour de thématiques comme les blogs, le développement durable, le militantisme sur Internet. Enfin, en 2009, la BnF a fait l'acquisition auprès d'Internet Archive des sites français collectés par la fondation entre 1996 et 2000, formant ainsi les « incunables » du Web français. Cette collection représente à elle seule 60 To de données, pour quelques 6 milliards de fichiers[85].

---

[84] GAME, Valérie, OURY, Clément, op. cit., p. 85.

[85] GAME, Valérie, OURY, Clément, op. cit., p. 70. ; GAME Valérie, ILLIEN Gildas, « Le dépôt légal d'Internet à la Bibliothèque nationale de France : cadre juridique, modèle de la collecte, évolution des métiers », *BBF*, 2006, n° 3, p. 84.

### Plan quadriennal de recherche : 2016-2019

La BnF jouit à présent d'une procédure d'archivage du Web bien rodée et longuement testée, améliorée. Le service du dépôt légal du Web s'est tourné depuis 2016 vers le service aux chercheurs et l'amélioration de ses outils d'analyse et au développement de corpus. L'indexation plein texte est, par exemple, expérimentée. Elle a été mise en place sur quelques corpus, comme celui des archives des attentats de Paris de janvier et novembre 2015. La BnF travaille avec les chercheurs pour cerner au mieux leurs besoins et leurs attentes et ainsi pouvoir améliorer ses propres services[86].

L'archivage du Web s'est rapidement développé dans le monde, même si les pays développés restent encore les principaux concernés. La France a pu profiter de la coopération internationale pour expérimenter l'archivage du Web français. Cette phase d'élaboration d'une politique documentaire et du perfectionnement de ses techniques et matériels informatiques ont permis de déployer efficacement la collecte large lors de l'application du décret de 2011. Depuis, plusieurs types de collecte ont lieu chaque année ainsi que des collectes exceptionnelles en lien avec l'actualité. On peut se demander comment la BnF, forte de tout cela, a-t-elle traité la question de la conservation des documents numériques et des archives du Web de manière pérenne et quels impacts l'archivage du Web a eu sur les métiers des bibliothèques ?

[86] MUSIANI Francesca, *et al.*, op. cit., p. 18.

# Bibliographie

## 1. Monographies

### Archiver le Web


CHAIMBAULT Thomas, *L'archivage du web*, Villeurbanne, Enssib, 2008.

HOOG Emmanuel, *L'Ina*, Paris, PUF, 2006.

MASANÈS Julien (dir.), *Web archiving*, Berlin, Springer, 2006.

MUSIANI Francesca, et al., *Qu'est-ce qu'une archive du Web ?*, Marseille, Open Edition Press, 2019.


### Le patrimoine numérique


BIBLIOTHEQUE NATIONALE D'AUSTRALIE, Directives pour la sauvegarde du patrimoine numérique, UNESCO, 2003.

KAVČIC-COLIĆ Alenka, *Archiving the Web - some legal aspects*, 68th IFLA Council and General Conference, 2002.

TRELEANI Matteo, Qu'est-ce que le patrimoine numérique ? Une sémiologie de la circulation des archives, Lormont, Le Bord de l'eau, 2017.


### Techniques de l'archivage et de la conservation numériques


BACHIMONT Bruno, Patrimoine et numérique. Technique et politique de la mémoire, Bry-sur-Marne, Ina, 2017.

BRÜGGER Niels, *Archiving websites: general considerations and strategies*, Aarhus, Center for Internet studies, 2005.

LUPOVICI (Catherine), *La conservation des publications électroniques et du dépôt légal*, UNESCO, 2007.

PAPY Fabrice, *Bibliothèques numérique. Interopérabilité et usages*, Paris, ISTE Éditions, 2015.

# 2. Articles

## Archiver le Web


AUBRY Sara, *et al.*, « Les archives de l'Internet : un nouveau service de la BnF », *Documentaliste-Sciences de l'Information,* 2008, vol. 45, n° 4, p. 12-14.

BRÜGGER Niels, « Web archiving - between past, present, and future », in CONSALVO Mia, ESS Charles (Eds.), *The Handbook of Internet Studies*, Oxford, Wiley-Blackwell, 2011, p. 24-42.

BRÜGGER Niels, « Web history and the web as a historical source », *Zeithistorische Forschungen*, 2012, n° 9, p. 316-325.

BRÜGGER Niels, « A brief outline of temporalities of the Web online and in Web archives », SCHAFER Valérie (dir.), *Temps et temporalités du Web*, Nanterre, Presses universitaires de Paris Nanterre, 2018, p. 57-74.

CHEVALIER Philippe, ILLIEN Gildas, « Les archives de l'internet. Une étude prospective sur les représentations et les attentes des utilisateurs potentiels », Rapport BnF, 2011.

CLAVERT Frédéric, OURY Clément, « Sommes-nous en train de perdre la mémoire ? Mémoire et archivage du web », in *THATCamp Paris 2012 : Non-actes de la non-conférence des humanités numériques*, Paris : Éditions de la Maison des sciences de l'homme, 2012.

KAHLE Brewster, « Preserving the Internet », *American Scientific*, n° 276, p. 82-83.

MUSSOU Claude, « Et le Web devint archive : enjeux et défis », *Le Temps des médias*, 2012, vol. 19, n° 2, p. 259-266.

NIU Jinfang, « An overview of web archiving », *D-Lib magazine*, 2012, vol. 18, n° 3-4.

SCHAFER Valérie, THIERRY Benjamin, « L'ogre et la toile. Le rendez-vous de l'histoire et des archives du Web », *Socio. La nouvelle revue des sciences sociales*, n° 4, 2015, p. 75-95.


## Le dépôt légal numérique


COHEN Évelyne, VERLAINE Julie, « Le dépôt légal de l'internet français à la Bibliothèque nationale de France », *Sociétés & Représentations*, 2013, vol. 1, n° 35, p. 209-218.

GAME Valérie, ILLIEN Gildas, « Le dépôt légal d'Internet à la Bibliothèque nationale de France : cadre juridique, modèle de la collecte, évolution des métiers », *BBF*, 2006, n° 3, p. 82-85.

GAME, Valérie, OURY, Clément, « Le dépôt légal de l'Internet à la BnF : adapter une mission patrimoniale à l'économie de l'immatériel », in INSTITUT NATIONAL DU PATRIMOINE, *Le patrimoine culturel au risque de l'immatériel : enjeux juridiques, culturels,*

*économiques : actes du colloque [colloque, Paris, INP, 3-4 avril 2008]*, Paris, L'Harmattan, 2010, p. 59-76.

GHARSALLAH Mehdi, « Archivage du Web français et dépôt légal des publications électroniques », *Documentalistes - Sciences de l'Information*, ADBS, 2004.

HUCHET Bernard, ILLIEN Gildas, OURY Clément, « Le temps des moissons : le dépôt légal du Web : vers la construction d'un patrimoine coopératif », *BIBLIOthèque(s)*, 2010, n.52, p. 28-31.

ILLIEN Gildas, « Le dépôt légal de l'internet en pratique : les moissonneurs du Web », *BBF*, 2008, n° 6, p. 20-27.

LASFARGUES France, OURY Clément, WENDLAND Bert, « Legal deposit of the French Web: harvesting strategies for a national domain », *International Web Archiving Workshop*, Aarhus, 2008.

## Le patrimoine numérique

BERMES Emmanuelle, OURY Clément, « Web 2.0 et mémoire : de la conversation à la conservation », *Documentaliste-Sciences de l'Information*, 2009, vol. 46, n° 1, p. 61-63.

MARCETTEAU-PAUL Agnès, « Le patrimoine, une valeur d'avenir ? », *BBF*, n° 5, 2004.

MELOT Michel, « Qu'est-ce qu'un objet patrimonial ? », *BBF*, n° 5, 2004.

MUSIANI Francesca, SCHAFER Valérie, « Patrimoine et patrimonialisation numériques », *Reset*, 2017, n° 6.

PALOQUE-BERGES Camille, SCHAFER Valérie, « Quand la communication devient patrimoine », in *Hermès*, n° 71, 2015, p. 255-261.

SCHAFER Valérie, MUSIANI Francesca, BORELLI Marguerite, « Le patrimoine culturel immatériel pour aider à penser le patrimoine activement numérique », *Patrimoine culturel immatériel et numérique*, M. Sévero, S. Cachat éd., Paris, L'Harmattan, 2017, p. 131-145.

WEGRZYN-WOLSKA Katarzyna, « Le document numérique dynamique : une « étoile filante » dans l'espace documentaire », in SAVARD Réjean (dir.), *Le numérique : impact sur le cycle de vie du document : actes du colloque [colloque, Montréal, EBSI-enssib, 13-15 octobre 2004]*, Villeurbanne, EBSI-enssib, 2004, p. 127-138.

VERRY Élisabeth (éd.), « Archives et Internet : contributions et témoignages », *La Gazette des archives*, 2007, n° 207/3.

### Collectes ciblées

OURY Clément, « Soixante millions de fichiers pour un scrutin. Les collections de sites politiques à la BnF », *Revue de la BnF*, 2012, vol. 40, n° 1, p. 84-90.

SCHAFER Valérie, « Le patrimoine nativement numérique des attentats en Europe : regards croisés », in *La Gazette des archives*, Paris, 2018 n° 250, p. 115-129.

### Coopération internationale

GOMES Daniel, MIRANDA João, COSTA Miguel, « A survey on Web archiving initiatives », *TPDL 2011. Proceedings of the 15th international conference on Theory and practice of digital libraries: research and advanced technology for digital libraries*, Berlin/Heidelberg/New York, Springer, 2011, p. 408-420.

ILLIEN Gildas, « L'archivage d'internet, un défi pour les décideurs et les bibliothécaires : scénarios d'organisation et d'évaluation ; l'expérience du consortium IIPC et de la BnF », *actes du 74e Congrès de l'IFLA, Fédération internationale des associations de bibliothécaires et d'institutions*, Québec, 2008.

ILLIEN Gildas, « Une histoire politique de l'archivage du web. Le consortium international pour la préservation de l'Internet », *BFF*, 2011, n° 2, p. 60-68.

### Techniques de l'archivage et de la conservation numériques

MERZEAU Louise, « Web en stock », *Cahiers de médiologie*, 2003, p. 158-167.

SCHAFER Valérie, MUSIANI Francesca, BORELLI Marguerite, « Negotiating the Web of the past. Web archiving, governance and STS », *French Journal for Media Research*, n° 6, numéro spécial *La toile négociée/Negotiating the Web*.

## 3. Sitographie

### Articles

#### Archiver le Web

AUBRY Sara, « Introducing Web archives as a new library service: the experience of the National Library of France », *LIBER Quaterly. The journal of the Association of European Research Libraries*, 2010 [En ligne : https://dspace.library.uu.nl/handle/1874/241599 consulté le 30/02/2019].

CARROLL Rory, « Brewster's trillions: Internet Archive strives to keep web history alive », *The Guardian*, 26 avril 2013 [En ligne : https://www.theguardian.com/technology/2013/apr/26/brewster-kahle-internet-archive consulté le 02/05/2019].

DE LA PORTE Xavier, « Brewster Kahle, Internet Archive : « le meilleur du web est déjà perdu » », *internetactu.net*, 2011 [En ligne : http://www.internetactu.net/2011/06/28/brewster-kahle-internet-archive-le-meilleur-du-web-est-deja-perdu/ consulté le 20/01/2019].

DE LA PORTE Xavier, "Archiver le Web, c'est être condamné aux vides", *La vie numérique*, France Culture, 2016 [En ligne : https://www.franceculture.fr/emissions/la-vie-numerique/archiver-le-web-cest-etre-condamne-aux-vides consulté le 07/03/2019].

KAHLE Brewster, « Internet Archive, le meilleur du web est déjà perdu », 2011 [En ligne : http://www.internetactu.net/2011/06/28/brewster-kahle-internet-archive-le-meilleur-du-web-est-deja-perdu/ consulté le 07/12/2018].

### Le dépôt légal numérique

JACQUOT Olivier, « Le web des années 1990 : 20 ans d'*Internet Archive* et 10 ans du dépôt légal du web en France », *Carnet de la recherche à la Bibliothèque nationale de France*, 19 novembre 2016 [En ligne : https://bnf.hypotheses.org/1309 consulté le 10/12/2018].

MERZEAU Louise, MUSSOU Claude, « L'expérience des ateliers du dépôt légal du Web de l'Ina », *Carnet de recherche WebCorpora*, 2017 [En ligne : https://webcorpora.hypotheses.org/302 consulté le 30/03/2019].

ROUX Pauline, « Le dépôt légal du Web : Des tweets aux e-books, la conservation 2.0 », *Monde du livre*, 22 décembre 2016 [En ligne : https://mondedulivre.hypotheses.org/5697 consulté le 20 décembre 2018].

### Collectes ciblées

BORELLI Margueritte, « Documenting the 2015 Paris attacks web archives: An interview with the Archive-It team », *Carnet de recherche Hypothèses ASAP*, BnF, Paris, 21 mars 2016 [en ligne : https://asap.hypotheses.org/125 consulté le 21/01/2019]

BORELLI Margueritte, SCHAFER Valérie, "Entretien autour des collectes d'urgence au moment des attentats de janvier et novembre 2015 avec Annick Le Follic, Chargée de collections numériques, au département de dépôt légal de la BnF", *Carnet de recherche Hypothèses ASAP*, BnF, Paris, 21 mars 2016 [En ligne : https://asap.hypotheses.org/168 consulté le 21/01/2019].

BORELLI Margueritte, SCHAFER Valérie, "Entretien autour des collectes d'urgence au moment des attentats de janvier et novembre 2015 avec Thomas Drugeon, responsable du DL Web

INA", *Carnet de recherche Hypothèses ASAP*, BnF, Paris, 21 mars 2016 [En ligne : https://asap.hypotheses.org/173 consulté le 21/01/2019].

### Coopération internationale

GEBEIL Sophie, « Pourquoi archiver le Web ? Les missions de l'IIPC », *Carnet de recherche Internet, histoire et mémoires,* 2014 [En ligne : https://madi.hypotheses.org/243 consulté le 12/02/2019].

### Le Web 2.0

BERGMAN Michael K., « The 'Deep' Web: surfacing hidden value », *BrightPlanet*, [En ligne : https://brightplanet.com/2012/06/the-deep-web-surfacing-hidden-value/ consulté le 04/01/2019].

LYMAN Peter, VARIAN Hal, « How much information ? », dans *School of information management and systems,* 2003 [ En ligne : http://groups.ischool.berkeley.edu/archive/how-much-info-2003/ consulté le 20 décembre 2018].

# Sites Web

## Les Institutions

Bibliothèque nationale de France : https://www.bnf.fr/fr/archives-de-linternet

Bibliothèque nationale de France : https://www.bnf.fr/fr/spar-systeme-de-preservation-et-darchivage-reparti

Internet Archive : https://archive.org/

Internet Archive (blog) : https://blog.archive.org/

International Internet Preservation Consortium : http://netpreserve.org/

Institut national de l'audiovisuel : https://www.ina.fr/

Unesco : https://fr.unesco.org/

### Les programmes d'archivage du Web

Archive-It : https://archive-it.org/

List of Web archiving initiatives : https://en.wikipedia.org/wiki/List_of_Web_archiving_initiatives

NEDLIB : https://cordis.europa.eu/project/rcn/43331/factsheet/fr

PANDORA : http://pandora.nla.gov.au/

PROMISE : https://promise.hypotheses.org/

### Etudes universitaires et outils

ASAP - Archives sauvegarde attentats Paris : https://asap.hypotheses.org/

BnF API et jeux de données : http://api.bnf.fr/liste-des-adresses-url-des-collectes-ciblees-du-web-francais-par-la-bnf

Carnet de la recherche à la Bibliothèque nationale de France : https://bnf.hypotheses.org/1105

Web Corpora : https://webcorpora.hypotheses.org/200#more-200

# Étude de cas : l'archivage du Web par la Bibliothèque nationale de France et son influence sur les métiers des bibliothèques

Cette étude de cas est l'occasion de revenir plus en détail sur la mission d'archivage du Web menée par la Bibliothèque nationale de France depuis 2011. Il y sera d'abord question des différents types de collecte employés et de leurs intérêts respectifs. Ensuite, nous étudierons les évolutions que l'archivage du Web a apportées aux métiers des bibliothèques.

## 1. Méthodologie

Cette étude de cas a été rendue possible grâce à la coopération de la BnF qui a fourni les documents et informations nécessaires. Afin de présenter au mieux les différentes collectes organisées et leurs caractéristiques, le choix a été fait d'analyser dans un premier temps deux collectes d'urgence menées en 2015 lors des attentats de Paris. Ces deux collectes présentent plusieurs avantages. D'abord, elles portent toutes les deux sur le même type d'événement qui a eu un impact national fort et des répercussions au-delà du territoire français. Ensuite, pouvoir mener des collectes d'urgence implique nécessairement d'avoir une procédure de collecte déjà bien définie et rodée en amont pour pouvoir être aussi réactif que possible dans des cas exceptionnels, tels que des attentats ou tout événement extraordinaire avec impact sur la société. La consultation des archives du Web français se fait par recherche d'adresses URL précises. Nous avons donc demandé au service du dépôt légal numérique de la BnF de nous fournir la liste des URL collectées en janvier et en novembre 2015. Ces listes Excel ont permis d'étudier les différences de choix et de traitement entre les deux collectes. Leur étude approfondie a permis d'extraire les types de pages Web collectées, la part d'aide internationale et les choix documentaires appliqués à ces collectes.

En complément, les entretiens menés par Marguerite Borelli et Valérie Schafer en 2016 autour de ces collectes à la BnF, à l'Ina et auprès d'Archive-It ont permis d'avoir des renseignements détaillés sur le déroulement de chacune de ces collectes. Nous avons donc choisi de ne pas mener d'entretien sur ce sujet. Pour compléter et mettre en lumière le fonctionnement courant des archives du Web au sein de la BnF, nous avons choisi de comparer ces collectes d'urgence aux deux autres types de collecte, à savoir les collectes courantes et les collectes ciblées.

Ensuite, pour actualiser les données recueillies dans nos lectures, nous avons mené un entretien au sein du service du dépôt légal numérique à la Bibliothèque nationale de France. Cet entretien a eu lieu le 25 avril 2019 et s'est déroulé selon un mode semi-directif. La première partie de cet entretien est une présentation libre du service du dépôt légal

numérique, où son histoire est retracée et son fonctionnement expliqué. La seconde partie est plus directive puisqu'elle s'appuie sur un questionnaire préparé en amont. Ces questions portent principalement sur les évolutions des métiers des bibliothèques autour de l'archivage du Web ainsi que sur des points plus techniques du service. Cet entretien très riche a permis d'actualiser les informations sur le fonctionnement du service ainsi que sur les méthodes de stockage des données archivées et leur pérennisation au sein de la Bibliothèque nationale de France.

# 2. Les procédures de collecte, étude au regard de deux collectes d'urgence

La BnF pratique deux types de collecte durant l'année. D'abord, la collecte large annuelle puis la collecte ciblée qui permet de choisir des thématiques particulières et ainsi compléter d'éventuelles lacunes de la collecte large. En parallèle de ces deux modèles, la BnF a instauré une procédure de collecte d'urgence pour refléter dans ses collections les événements marquants de la société française en archivant les réactions suscitées sur le Web. Ces collectes d'urgence présentent quelques particularités par rapport aux procédures programmées.

## 2.1. Les collectes d'urgence autour des attentats de Paris en 2015

L'année 2015 a été marquée par plusieurs attentats, dont deux en France, à Paris. Le premier, le 7 janvier, a touché la rédaction de l'hebdomadaire *Charlie Hebdo*. Le second a touché plusieurs rues de la capitale le 13 novembre. La presse a couvert ces événements et les internautes du monde entier ont également réagi, notamment sur les réseaux sociaux. Des particuliers ont même commencé à collecter et sauvegarder de leur côté ces réactions, comme Nick Ruest, chercheur à l'université de York. Au sein de la BnF, comme au sein de l'Ina, des collectes d'urgence ont été lancées pour couvrir ces événements et en conserver une trace détaillée dans les archives. De ces deux collectes résultent le traitement de plusieurs millions de données[87].

[87] SCHAFER Valérie, op. cit., p.117.

### 2.1.1. Contextualisation

En temps que dépositaire du dépôt légal du Web, la BnF a pour obligation de collecter une fois par an le domaine français. Cependant, cette collecte a toujours semblé insuffisante en soi au département du dépôt légal numérique qui a cherché à la compléter. La loi DADVSI a l'avantage d'être assez évasive sur les modalités de ces collectes, donnant ainsi une marge de manœuvre confortable pour la BnF et l'Ina. Ainsi, la BnF s'est mise à réaliser des collectes plus ciblées dès 2007 autour des présidentielles. La BnF a bénéficié en 2010 d'un accroissement du nombre de ses serveurs destinés à l'archivage du Web. À la tête de 50 serveurs, le service du dépôt légal numérique dispose en permanence de matériel libre et opérationnel pour lancer une collecte dans la journée. Pour encadrer cette procédure d'urgence envisagée, une charte documentaire a été créée, en concertation avec les collaborateurs du réseau. Ainsi, la BnF a défini « *[qu'] une urgence en termes de collecte du Web peut être liée à un événement éphémère ou liée à la disparition prochaine d'un site Web*[88] ».

Suite aux expérimentations autour des collectes de 2007, la BnF a également mis en place en 2010 deux collectes automatiques intitulées « Actualités » et « Presse payante ». Elles permettent de collecter tous les matins à 10 h la page d'accueil et un clic de 100 titres de presse français accessibles gratuitement et de faire la même chose à 14 h pour 25 titres payants, majoritairement régionaux. Ces deux collectes autour de la presse assurent une part essentielle des collectes d'urgence puisqu'elles assurent la veille et la capture des articles de presse en plus de fonctionner également le week-end, alors que la Bibliothèque est fermée. Le reste de la collecte peut ainsi se concentrer sur la capture des réseaux sociaux numériques tels que Twitter[89].

#### Disponibilité des moyens humains et techniques

Les collectes d'urgence dépendent entièrement de l'intervention manuelle. Ainsi, le moment où se déroule l'événement joue un rôle majeur dans le résultat final de la collecte. Dans le cas des attentats contre *Charlie Hebdo*, les faits se sont produits un mercredi, tout le service a pu réagir rapidement. Mais les attentats de novembre ont eu lieu un vendredi soir. La BnF est fermée le samedi et le dimanche et l'accès aux outils et serveurs du dépôt légal numérique n'est possible que dans l'enceinte de la bibliothèque, pour des raisons de sécurité.

---

[88] BORELLI Marguerite, SCHAFER Valérie, "Entretien autour des collectes d'urgence au moment des attentats de janvier et novembre 2015 avec Annick Le Follic, Chargée de collections numériques, au département de dépôt légal de la BnF", *Carnet de recherche Hypothèses ASAP*, BnF, Paris, 21 mars 2016. https://asap.hypotheses.org/168

[89] Ibid.

Il a donc fallu attendre le lundi matin, soit plus de 48 h après les faits pour que la collecte d'urgence soit lancée. Dans cette situation, on comprend l'importance du rôle des collectes automatiques autour de la presse. Cela a permis de collecter une partie des premières réactions autour de l'attentat[90].

Enfin, il faut prendre en compte les autres activités du service du dépôt légal numérique. Les équipes ne sont pas toujours pleinement disponibles pour se consacrer à temps plein à une collecte d'urgence. Les attentats contre *Charlie Hebdo* ont eu lieu lors d'une période creuse au sein du service. Cela a permis de mobiliser deux personnes à temps-plein durant un mois pour mener à bien cette collecte, puisqu'elles ont pu repousser leurs autres activités. L'attentat de novembre a eu lieu durant une période moins favorable pour le service. Plusieurs collectes étaient en cours, d'une part la collecte large annuelle, mais aussi des collectes ciblées autour des élections régionales, de la COP21 ainsi qu'autour des réfugiés. En plus de mobiliser l'ensemble de l'équipe, ces collectes ont aussi réduit le nombre de serveurs disponibles pour une nouvelle collecte immédiate. La collecte d'urgence a donc été réduite à une semaine et la masse de données collectées considérablement réduite, par les limites techniques et aussi par choix délibéré de collecter moins, mais mieux. Pour rappel, le dépôt légal numérique vise la représentativité et non l'exhaustivité, cela vaut plus que jamais pour les collectes d'urgence.

### 2.1.2. La collecte d'urgence appliquée

Une fois la décision prise en réunion de démarrer une collecte d'urgence pour les attentats, il a fallu faire une première sélection d'URL à collecter pour les envoyer au DSI, département des systèmes d'information, qui a alors lancé Heritrix avec ces instructions. Pour les attentats contre *Charlie Hebdo*, la couverture a été bien plus large que pour ceux de novembre. Pour les réseaux sociaux numériques, la collecte a été plus simple que celle de novembre. L'essentiel des *hashtags* étaient concentrés autour de #jesuischarlie et de ses variantes. Pour novembre, les mots-clés étaient plus diversifiés. En dehors de tweets particuliers présélectionnés, la BnF a réalisé 4 captures par jour du fil Twitter. Chaque capture représente un affichage, sans déroulement du fil, soit 20 tweets par capture. La BnF a aussi collecté une fois par jour quelques comptes et groupes Facebook[91]. Si l'Ina utilise l'API de Twitter pour ses collectes, la BnF utilise Heritrix y compris pour les réseaux sociaux. Contrairement à Twitter qui est entièrement public, les pages Facebook sont généralement privées, donc plus difficiles à collecter. Pour compléter ses propres listes d'URL, la BnF a fait

[90] Ibid.

[91] Ibid.

appel à son réseau de correspondants au sein de la BnF et dans les pôles associés pour qu'ils fassent eux aussi remonter les pages qui leur semblaient pertinentes. Cela a permis d'élargir progressivement le périmètre de collecte, en ajoutant par exemple des sites de dessinateurs de presse[92].

Lors de ces deux collectes, la BnF a appliqué le même principe de représentativité que pour ses collectes larges. Cela veut dire qu'aucune censure n'a été appliquée quant au contenu des tweets et pages Web archivées. Sur le principe du dépôt légal, la BnF a défini que « *l'application de critères normatifs ou moraux n'est pas [son] rôle. [Sa] démarche et de capturer la variété : des points de vue, formes, publications techniques, émetteurs, etc.*[93] ». Ainsi, il n'est pas exclu que des contenus « *jugés illicites par la police après leur publication*[94] » fassent partie des collections issues de ces collectes. Le cas ne s'est à ce jour pas présenté, mais si les autorités exigent le retrait de ces contenus, ils ne seront évidemment pas détruits. Ils seront seulement retirés de la consultation.

### Contributions étrangères

En parallèle de sa propre collecte d'urgence, la BnF a bénéficié du soutien de plusieurs institutions étrangères, principalement pour la collecte de janvier. Il faut rappeler que la BnF fait partie de l'IIPC et bénéficie donc d'un important réseau de bibliothèques à travers le monde. Très tôt dans la collecte, la BnF a reçu des messages de sympathie de plusieurs membres de l'IIPC. Ces derniers ont envoyé des listes d'URL qu'ils avaient constitués. Ces listes sont pour une grande part composées de sites relevant du domaine national de chaque institution, ainsi que des domaines plus généraux ou encore parfois relevant du .fr. Archive-It, membre de l'IIPC et branche d'Internet Archive a rapidement offert sa collaboration à la collecte. En plus de proposer des URL, Archive-It a reçu la copie des URL sélectionnées par la BnF et les autres institutions, pour organiser une seconde archive et répartir le travail de veille. La collection constituée par Archive-It est identique à celle réalisée par la BnF. Seulement, Archive-It permet l'accès public à ses collections tandis que le cadre légal impose une consultation sur place pour la BnF[95].

---

[92] Ibid.

[93] Ibid.

[94] Ibid.

[95] Ibid. ; BORELLI Marguerite, « Documenting the 2015 Paris attacks web archives: An interview with the Archive-It team », *Carnet de recherche Hypothèses ASAP*, BnF, Paris, 21 mars 2016. https://asap.hypotheses.org/125

La coopération internationale entre programmes d'archivage du Web se fait donc aussi lors de période de collectes exceptionnelles. La BnF a pu bénéficier de ce soutien en janvier 2015, avec la participation de 15 institutions. En novembre, la participation internationale s'est limitée à l'envoi des URL sélectionnées par la BnF à Archive-It pour créer une collection de leur côté. Cette différence de traitement s'explique notamment par l'écart de situation entre le temps disponible et le souhait de collecter des données plus qualitatives, faute d'être quantitatives. La BnF avait elle aussi déjà proposé son aide à une autre institution de l'IIPC pour une collecte étrangère. En 2011, elle avait contribué à la collecte sur Vaclav Havel organisée par la Bibliothèque nationale de République Tchèque[96].

### Analyse des deux collectes

L'analyse des listes d'URL des collectes de janvier et novembre 2015 fournies par la BnF dans le cadre de cette étude met en lumière la différence de traitement des deux événements. Alors que la collecte de janvier représente quelque 4 216 adresses URL, la collecte de novembre se compose de 66 adresses URL.

Les URL sélectionnées par la BnF représentent 41 % de la collecte de janvier. Les 59 % restant ont été signalés par des institutions étrangères. On en compte 15, pour la plupart situées en Europe. Cependant, comme le graphique suivant le montre, toutes n'ont pas envoyé une quantité équivalente d'URL.

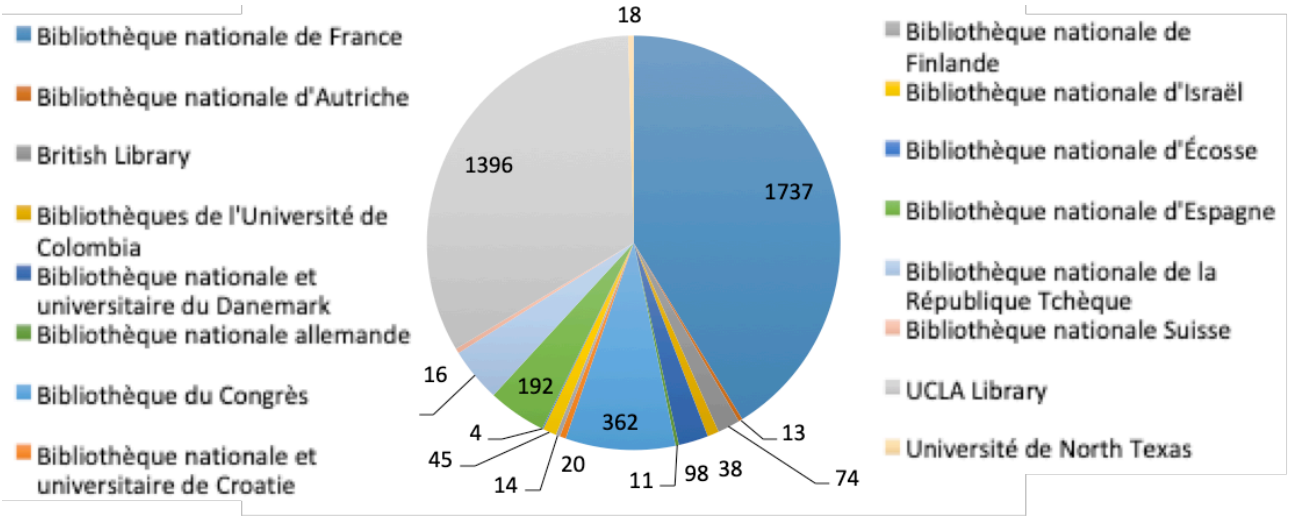



Figure 1 : Répartition des adresses URL envoyées par des institutions.

[96] BORELLI Marguerite, SCHAFER Valérie, "Entretien autour des collectes d'urgence au moment des attentats de janvier et novembre 2015 avec Annick Le Follic, Chargée de collections numériques, au département de dépôt légal de la BnF". https://asap.hypotheses.org/168

La plupart ont envoyé moins d'une centaine d'adresses quand d'autres, comme l'UCLA Library en a proposé plus de 1 300. Cette disparité s'explique notamment par le choix du périmètre de chaque institution. Par exemple, la Bibliothèque nationale d'Écosse n'a sélectionné que des adresses de sites officiels du domaine .org édités en Écosse. De même, la Bibliothèque nationale d'Autriche a sélectionné des URL en .at et quelques .com de sites de villes autrichiennes comme Salzbourg. A contrario, si la Bibliothèque du Congrès a pu proposer 362 URL, c'est parce que son périmètre était bien plus étendu. Elle a ainsi envoyé des adresses du domaine américain, mais aussi des sites en .uk.co, en .fr ou encore en .de. Cette multiplicité des participations et des périmètres a permis à la BnF d'obtenir une couverture internationale des réactions autour des attentats de janvier. Cette aide illustre parfaitement la place accordée à la coopération internationale autour de l'archivage du Web et du véritable réseau qui s'est construit au sein de l'IIPC.

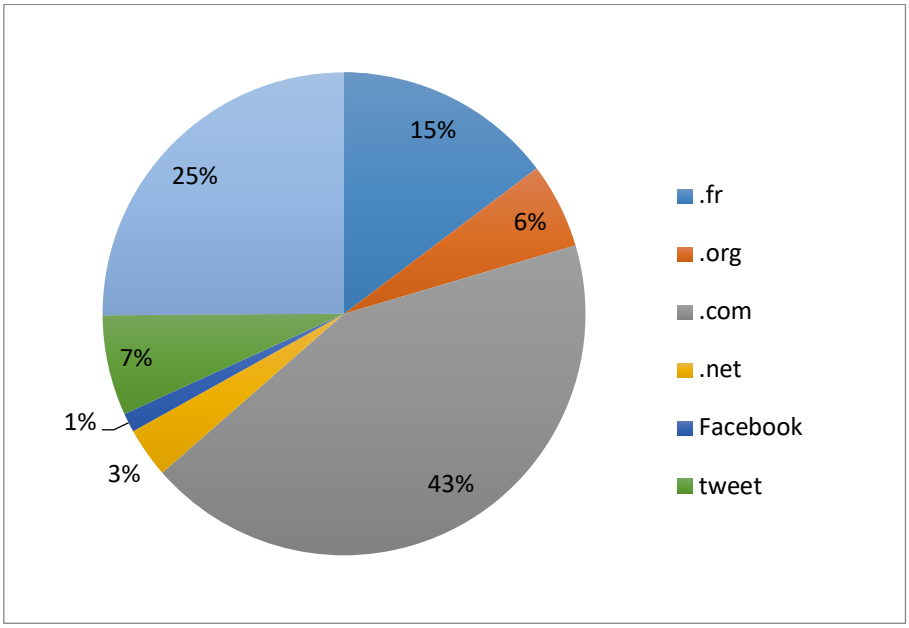



Figure 2 : Répartition des noms de domaine dans la collecte d'urgence autour des attentats de Paris en janvier 2015.

Cette collecte d'urgence montre également la faible part des domaines nationaux par rapport aux domaines en .com ou .net. En analysant ce graphique, on constate que le domaine .com représente 43 % des 4 216 URL collectées. Il représente ainsi la plus grande part de ces URL. Cependant, si l'on additionne la part du .fr aux autres domaines nationaux, on obtient 40 % du total. Sans l'apport d'institutions extérieures, un quart de ces adresses URL n'aurait pas pu être collecté et archivé par la BnF. Le choix a été fait d'isoler de ce graphique les collectes de pages Facebook et de tweets pour mettre en lumière la part non-négligeable prise par les réseaux sociaux. On obtient 7 % pour Twitter et seulement 1 % pour Facebook. Cette différence s'explique par la difficulté de collecter les pages Facebook du fait de leur caractère

privé. Cela pose malgré tout un problème de représentativité puisque les Français sont bien plus présents sur Facebook que sur Twitter[97].

La collecte sur Twitter a été faite en suivant des *hashtags* et des comptes bien précis. La BnF a ainsi établi la liste de 10 *hashtags* à collecter : #CharlieHebdo, #jenesuispascharlie, #jesuischarlie, #JeSuisCharlie, #noussommescharlies, #NousSommesCharlie, #TousALaMarcheDu11Janvier, #LaMarcheDu11Janvier, #LaFranceEstCharlie et #ratp. Les différentes graphies ont été prises en compte pour ne pas rater une partie des tendances majeures. En complément, 59 comptes Twitter ont été suivis activement, parmi lesquels des officiels tels que François Hollande, alors président de la République, le ministre de l'Intérieur Bernard Cazeneuve, la maire de Paris Anne Hidalgo, ou encore des comptes de presse comme *Le Monde*, *Le Point*, *Euronews* et bien évidemment le compte de *Charlie Hebdo*.

Les attentats de novembre 2015 ont abouti à une collecte bien plus restreinte. D'une part pour les raisons techniques et humaines précédemment évoquées et d'autre part pour ne pas reproduire la surabondance de la collecte de janvier. En effet, parmi les 4 216 URL archivées, la BnF ne les estime pas toutes pertinentes et quelques redondances apparaissent. La collecte de novembre est donc composée de seulement 66 URL. Cette collecte a été complétée en parallèle avec les deux collectes quotidiennes autour de l'actualité, les sites de presse n'ont donc pas été sélectionnés directement pour la collecte d'urgence afin d'éviter les doublons.

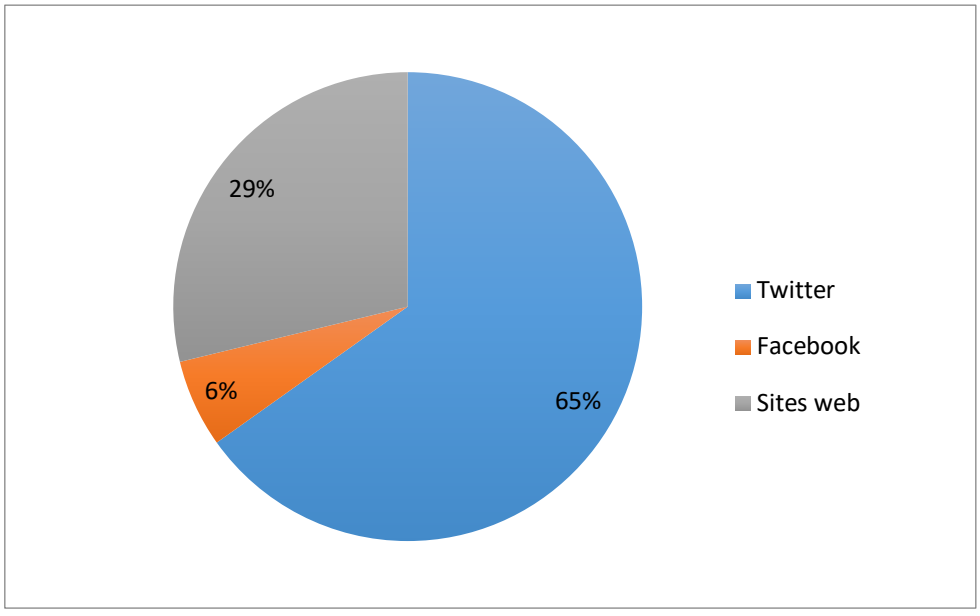



Figure 3 : Répartition des adresses URL collectées autour des attentats de Paris en novembre 2015.

[97] MUSIANI Francesca, et al., op. cit., p. 42.

L'analyse de la répartition des URL collectées montre une prépondérance de Twitter alors que seulement 4 pages Facebook sont sélectionnées. Plusieurs comptes officiels font partie de cette liste, comme Anne Hidalgo, le compte du ministère de la Défense, la gendarmerie ou encore les pompiers de Paris. S'ajoutent à ces comptes 24 *hashtags*, parmi lesquels on trouve #attackparis, #Bataclan, #CRS, #FluctuatNecMergitur, #prayforparis ou #tousaubistrot.

Les 29 % restants sont constitués de sites Web, tous français. On compte 11 sites officiels du gouvernement. Il s'agit principalement des ministères de la Défense, de la Sécurité intérieure, de la Santé ou bien du site de l'Élysée. Les 8 sites restants sont des sites d'organisation comme la Croix-Rouge, des sites d'information, le site du Bataclan et la page Wikipédia « Attentats du 13 novembre 2015 en France ».

Ces deux collectes d'urgence montrent une évolution dans l'ampleur du périmètre de collecte. Alors que pour janvier, l'idée était d'être aussi exhaustif que possible, ce qui a généré des millions de données archivées, la collecte de novembre a été plus restreinte. On a alors plutôt visé la représentativité, comme pour les collectes courantes du dépôt légal, ce qui a généré un corpus moins volumineux mais pas moins qualitatif.

### 2.1.3. Les limites des collectes d'urgence

Les collectes d'urgence ont été créées pour compléter les collectes courantes et répondre à l'actualité. En ce sens, elles ont l'avantage de permettre de constituer des corpus autour d'événements particuliers. Cependant, les collectes d'urgences rencontrent quelques limites.

#### Des limites techniques et humaines

La réalisation d'une collecte d'urgence est avant tout conditionnée par les moyens techniques et humains disponibles à ce moment. Des serveurs doivent être disponibles pour pouvoir lancer Heritrix. Il faut aussi avoir une équipe disponible pour se consacrer à tout le processus de recherche et sélection des URL à collecter pour pouvoir les transmettre au robot. À ces prérequis s'ajoute la nature même du dépôt légal du Web. Là où le dépôt légal des livres imprimés vise l'exhaustivité, le dépôt légal du Web cherche la représentativité, ne pouvant couvrir les 8 millions, estimés, de noms de domaine du Web français[98]. Ainsi, la durée d'une collecte d'urgence dépend de la disponibilité des équipes et des serveurs. Leur lancement est aussi conditionné par les jours d'ouverture de la bibliothèque. Dans le cas des attentats de novembre, il a fallu attendre le lundi matin pour lancer la collecte d'urgence. L'Ina a pu la

[98] Annexe 3 : entretien à la BnF.

démarrer dès le vendredi soir grâce à un hasard du calendrier de maintenance. Une opération de maintenance sur le réseau électrique de la salle des serveurs était programmée pour ce jour-là, ce qui a nécessité de délocaliser certaines captations pour continuer d'assurer la captation directe des flux radio et TV sur Twitter. L'équipe d'astreinte a donc bénéficié d'un accès à distance aux outils de collecte et a pu lancer une procédure d'urgence quelques heures seulement après les attentats[99].

Les collectes d'urgence moissonnent parmi les réseaux sociaux numériques, car ils concentrent les réactions du monde entier, aussi bien d'acteurs politiques ou reconnus que de n'importe quel internaute anonyme. Ces réseaux sociaux ne sont pas toujours aisés à capturer. L'utilisation de protocoles de sécurité par les sites, comme les captcha ou le HTTPS nécessitent une intervention manuelle. Il faut donner une instruction manuelle à Heritrix pour s'assurer qu'il capture ce qui a été demandé. Cette intervention humaine directe ralentit le processus alors que le principe même de la collecte d'urgence réside dans sa réactivité. Le travail de veille et de maintenance est bien plus important que lors des collectes courantes[100].

### Des limites dans l'estimation de la valeur documentaire

La sélection des URL à collecter est un choix en soi des équipes, mais ces dernières ne peuvent suivre et appréhender l'ensemble des phénomènes autour d'un événement. Il n'est donc pas exclu qu'un mouvement ait échappé à la collecte. Pour la collecte courante, il s'agit d'un échantillonnage représentatif du Web à visée large. Pour les collectes d'urgence, cette question de représentativité est d'autant plus importante qu'elles constituent des collections ciblées qui permettront de documenter ces événements. Il est impossible d'estimer l'étendue des éventuelles lacunes dans ces collections. Ce sera aux chercheurs qui travailleront à partir de ces corpus à l'avenir de juger de la valeur documentaire de ces collections[101].

---

[99] BORELLI Marguerite, SCHAFER Valérie, "Entretien autour des collectes d'urgence au moment des attentats de janvier et novembre 2015 avec Thomas Drugeon, responsable du DL Web INA", *Carnet de recherche Hypothèses ASAP*, BnF, Paris, 21 mars 2016. https://asap.hypotheses.org/173

[100] BORELLI Marguerite, SCHAFER Valérie, "Entretien autour des collectes d'urgence au moment des attentats de janvier et novembre 2015 avec Annick Le Follic, Chargée de collections numériques, au département de dépôt légal de la BnF". https://asap.hypotheses.org/168

[101] BORELLI Marguerite, SCHAFER Valérie, "Entretien autour des collectes d'urgence au moment des attentats de janvier et novembre 2015 avec Thomas Drugeon, responsable du DL Web INA", *Carnet de recherche Hypothèses ASAP*, BnF, Paris, 21 mars 2016. https://asap.hypotheses.org/173

## 2.2. Les différences avec les autres collectes

La Bibliothèque nationale de France réalise couramment des collectes larges et des collectes ciblées. La collecte large permet d'assurer la mission de dépôt légal du Web annuellement. Les collectes ciblées assurent quant à elles une plus grande profondeur de traitement sur des thématiques précises.

### 2.2.1. La collecte large

La collecte large assure le moissonnage de l'ensemble du domaine français en .fr ainsi qu'un échantillon de sites hébergés sous un autre domaine, mais édités en France. Avec ce mode de collecte, la BnF estime couvrir environ 60 % du Web français. Une fois la liste des URL à moissonner établie, elle est envoyée au DSI. Ce dernier la transmet au *crawler* Heritrix et règle tous les paramétrages de la collecte. Pour chaque collecte, un budget est défini. Ce budget désigne le nombre maximal d'URL que peut collecter le *crawler* sur chaque site. Ainsi, la liste d'URL transmise initialement à Heritrix ne constitue que la base de la collecte. Pour chaque URL initiale, il va cliquer de lien en lien jusqu'à atteindre le budget puis passer à l'URL suivante. Pour une adresse URL initiale, qui représente un site Web, une multitude d'autres URL viennent s'ajouter. Heritrix simule le comportement d'un internaute. Il navigue de lien en lien depuis la page d'accueil du site jusqu'à ce qu'il atteigne le budget. La collecte large est un traitement de masse, aucune volonté documentaire n'y est associée. Cela permet d'assurer la mission de représentativité du Web français tout en minimisant l'intervention humaine sur le moissonnage en lui-même. Les équipes du dépôt légal du Web et du DSI n'interviennent que pour assurer la maintenance technique, la surveillance et le contrôle de la collecte[102].

Ainsi, pour la collecte large de 2018, Heritrix a mené son moissonnage à partir d'une liste de 4 500 000 noms de domaine. Le budget, quant à lui, a été fixé à 2 500 URL. Cette limitation est suffisamment grande pour permettre de collecter dans leur intégralité 97 % des sites sélectionnés au départ. Pour les 3 % restants, il s'agit de sites de grande envergure, comme la partie française de Wikipédia par exemple, qui représente à elle seule autour de 4 millions d'URL[103]. Une collecte large dure en moyenne entre deux et trois semaines[104]. La collecte de 2018 représente à elle seule un volume d'un peu plus de 100 To.

---

[102] Annexe 3 : entretien à la BnF.

[103] Ibid.

[104] ROUX Pauline, « Le dépôt légal du Web : Des tweets aux e-books, la conservation 2.0 », *Monde du livre*, 22 décembre 2016. https://mondedulivre.hypotheses.org/5697

### Quelques difficultés

Comme pour les collectes d'urgence, les protocoles de sécurité représentent une des principales difficultés rencontrées par l'archivage du Web. Lorsqu'un site est protégé par un protocole d'authentification, il faut qu'Heritrix s'identifie pour entrer dans le site et le moissonner. La loi a prévu ce cas de figure et oblige le producteur du site de fournir à la BnF un compte utilisateur pour pouvoir mener à bien la collecte. Heritrix a également été conçu pour pouvoir gérer une authentification automatique. En pratique, cela ne fonctionne pas aussi bien que prévu. Les protocoles de sécurité sont souvent plus complexes qu'une simple demande d'identifiant et de mot de passe. Généralement, le site crée une session spécifique pour chaque utilisateur. Un moteur de recherche classique mémorise l'identifiant de cette session et la réutilise à chaque nouvelle requête sur le site. Heritrix n'est pas capable de gérer cette fonction, il faudrait donc relancer manuellement l'authentification pour chaque page. Une telle intervention allongerait le traitement de la collecte de plusieurs jours, voire de plusieurs semaines. Face à ce problème, la BnF a choisi, hormis pour les sites de presse, d'ignorer ces sites pour les collectes larges.

Les bases de données constituent également un problème pour le moissonnage. Le plus souvent, les bases de données fonctionnent sur la base d'un référencement par mot-clé. L'utilisateur tape sa recherche sous forme de mots-clés et reçoit des documents correspondants. L'essentiel des documents n'est donc pas accessible via des liens directs, mais par cette requête. Or, Heritrix n'est pas capable de gérer cette fonctionnalité. Il ne peut pas « *taper tous les mots-clés imaginables en fonction du domaine, de la discipline à laquelle se rattache le site*[105] » pour obtenir des réponses. Ce contenu échappe donc totalement à la collecte. Une réflexion est en cours autour de ce problème à la BnF. La solution envisagée serait de créer un dépôt volontaire. Les éditeurs de base de données pourraient ainsi déposer directement leurs contenus, ce qui assurerait l'archivage complet par la BnF[106].

## 2.2.2. La collecte ciblée

Pour compléter les collectes larges, la BnF organise plusieurs collectes ciblées dans l'année, en partenariat avec ses correspondants au sein de la bibliothèque, mais aussi dans les pôles associés. Ces collectes sont aussi le moyen de refléter les nouvelles formes d'édition du Web, comme par exemple les blogs au début des années 2000[107]. Contrairement à la collecte

[105] Annexe 3 : entretien à la BnF.

[106] Ibid.

[107] CHAIMBAULT Thomas, op. cit., p. 35.

large, la collecte ciblée est caractérisée par un but documentaire. On peut l'apparenter à un zoom sur une discipline ou une thématique. Les collectes de presse sont des collectes ciblées.

Le service du dépôt légal numérique demande à ses correspondants de lui fournir une liste de sites Web français qui présentent un intérêt pour leur domaine. Ces correspondants sont des bibliothécaires des différents départements de la BnF ainsi que des bibliothèques en charge du dépôt légal imprimeur. Le service du dépôt légal numérique donne en amont quelques axes pour encadrer ses propositions, comme l'obligation de choisir des sites français pour rester dans le cadre du dépôt légal ou de bannir tout contenu illégal, comme l'incitation à la haine raciale, à la violence, etc. Cependant, aucune directive documentaire n'est donnée. Le service considère que chaque correspondant suit la politique documentaire de son service.

Les collectes ciblées jouissent d'un budget bien plus étendu que les collectes larges. Il est généralement compris entre 45 000 et 200 000 URL. Ce type de collecte permet de moissonner plus en profondeur. Par exemple dans le cas de la partie française de Wikipédia, certaines sections ont été sélectionnées pour des collectes ciblées alors qu'elles ont échappé au *crawler* lors de la collecte large. Il en va de même pour les bases de données et plateformes comme Cairn.info ou Open Edition. La fréquence de ces collectes est également plus élevée, allant du semestriel au quotidien pour Twitter et la presse. Au total, chaque année, les collectes ciblées représentent un volume compris entre 30 et 40 To pour environ 30 000 URL[108]. Il s'agit d'un échantillonnage qualitatif en opposition à la démarche quantitative de la collecte large.

À la fin de la collecte large de 2018, le volume total des archives du Web de la BnF a dépassé le pétaoctet. Si ces archives sont constituées de données allant de 1996 à nos jours, l'essentiel du volume a commencé à augmenter depuis 2010. Autour de 2005, une collecte large représentait environ 5 To tandis qu'aujourd'hui, on se situe autour des 100 To[109]. Cette différence de volumétrie s'explique d'une part par le budget qui a considérablement augmenté et d'autre part par l'expansion perpétuelle du Web français.

## 2.3. Stockage et pérennisation

Lorsqu'on évoque l'archivage des documents numériques et par extension des archives du Web, la question des méthodes de stockage, des moyens techniques et de la pérennisation de ces collections revient souvent au cœur des réflexions. La Bibliothèque nationale de France se place comme précurseur sur ces aspects.

[108] Annexe 3 : entretien à la BnF.

[109] Ibid.

### 2.3.1. SPAR : Système de Préservation et d'Archivage Réparti

Le Système de Préservation et d'Archivage Réparti, SPAR, a été mis en œuvre par la BnF en 2010. L'objectif de SPAR est de préserver les données numériques de la BnF. Il s'agit d'un ensemble de logiciels, de serveurs et de baies de stockage, les Petabox, qui tirent leur nom de leur capacité de stockage d'un pétaoctet. Ce système est réparti parce qu'il « *permet la gestion de plusieurs copies des documents, sur plusieurs sites, afin de se prémunir contre les pertes et destructions*[110] ». Cette répartition permet de sectoriser les baies de stockage[111] et de restreindre les accès selon les besoins. Ainsi, les données disponibles à la consultation du public sont des copies des données archivées et destinées à la conservation.

Grâce à une veille et des migrations préventives, on s'assure que chaque donnée reste « *lisible, compréhensible et réutilisable sur le long terme, même si l'environnement technique et humain dans lequel ces documents ont été produits change*[112] ». Chaque fichier versé dans SPAR est soumis à un contrôle de son format. Si celui-ci ne permet pas d'assurer une pérennité dans la conservation, il est refusé. Il faut alors le convertir. Dans certain cas, le fichier est stocké malgré tout s'il n'y a pas de solution satisfaisante pour l'instant. Dans d'autre cas, on choisit de verser le fichier tel quel ainsi que sa version convertie, car celle-ci risque de subir des pertes ou des modifications de données lors du processus. Cela permet d'avoir le fichier original pour conserver, au moins pour un temps, l'intégrité des données[113].

SPAR a été conçu sur la base du protocole OAIS, *Open Archival Information System*. Il s'agit d'une norme internationale pour la préservation des données numériques. L'OAIS est utilisé par les archives, les bibliothèques et les musées notamment. Ce protocole a été choisi pour permettre l'interopérabilité de SPAR avec les systèmes de conservation et d'archivage d'autres institutions[114].

En plus d'y stocker le fruit des collectes d'archivage et les documents d'autres départements de la BnF, comme par exemple ceux de la bibliothèque numérique Gallica, ou des documents administratifs, on verse progressivement les anciennes données. Ainsi, les collectes de 1996 à 2000 qui avaient été achetées à Internet Archive sont en cours de versement dans SPAR, tout comme celles réalisées par la BnF avant 2010[115].

---

[110] https://www.bnf.fr/fr/spar-systeme-de-preservation-et-darchivage-reparti

[111] Annexe 4 : schéma fonctionnel SPAR.

[112] https://www.bnf.fr/fr/spar-systeme-de-preservation-et-darchivage-reparti

[113] Annexe 3 : entretien à la BnF.

[114] Ibid. ; https://www.bnf.fr/fr/spar-systeme-de-preservation-et-darchivage-reparti

[115] Annexe 3 : entretien à la BnF.

SPAR permet donc de protéger l'intégrité des données, de les retrouver grâce aux métadonnées et d'assurer une veille préventive pour garantir un archivage pérenne.

### 2.3.2. Format de fichier et serveurs

Pour archiver ses données, la BnF utilise le format WARC développé au sein de l'IIPC. Ce format est propre à l'archivage du Web et permet de stocker et de manipuler d'importants volumes de données. Le format WARC peut être comparé aux boîtes de conservation en archive. Un fichier WARC peut contenir jusqu'à 1 Go de données. On verse donc dans un ficher WARC les données et métadonnées collectées au fur et à mesure du moissonnage. Une fois la limite des 1 Go atteinte, on ouvre un nouveau fichier WARC. Contrairement aux boîtes utilisées pour les archives non-numériques, l'ordre de classement et d'arrivée des données n'a pas d'importance. Les métadonnées attachées à chaque donnée permettent de retrouver les fichiers souhaités[116].

Ensuite, pour éviter tout risque lié à l'obsolescence matérielle, la BnF procède régulièrement au changement de ses serveurs de stockage. Alors que la durée de vie garantie par les fabricants est autour de 9 à 10 ans, la BnF renouvelle son parc tous les 5 ans environ. Cela permet de limiter les risques de *crash disk*, mais cela a aussi un intérêt pour les performances des outils. Les ressources disponibles pour chaque serveur sont atteintes bien avant l'usure physique du serveur, ce qui crée des ralentissements dans le fonctionnement des outils. Ce renouvellement tous les 5 ans assure un fonctionnement optimal des outils et des services[117].

# 3. Quelles évolutions pour les métiers des bibliothèques avec l'archivage du Web ?

L'archivage du Web a apporté son lot de nouveautés au sein de la Bibliothèque nationale de France. La création du dépôt légal du Web a abouti à l'ouverture du service du dépôt légal numérique. Cette nouvelle mission a amené à repenser en partie les métiers des bibliothèques autour du numérique ainsi que le statut même de ces nouvelles collections du Web. La réflexion est toujours en cours pour faire évoluer encore aujourd'hui certains de ces aspects.

---

[116] GAME, Valérie, OURY, Clément, op. cit., p. 72. ; ILLIEN Gildas, « Le dépôt légal de l'internet en pratique : les moissonneurs du Web », p. 23.

[117] Annexe 3 : entretien à la BnF.

### 3.1.1. Des « archives » menées par des bibliothécaires

Lorsqu'on parle d'archives du Web, on peut s'interroger sur l'utilisation du mot « archives ». Les archives du Web sont-elles réellement des archives au sens strict du terme ? Si l'on prend la définition d'archives, il s'agit d'un « *ensemble de documents hors d'usage courant, rassemblés, répertoriés et conservés pour servir l'histoire d'une collectivité ou d'un individu*[118] ». Les archives du Web rassemblent, répertorient et conservent des ensembles de documents et données numériques pour servir l'histoire à différents degrés. Alors qu'en est-il du critère hors d'usage courant ? Ces archives répondent à la nécessité de conserver une trace de la production de savoirs en ligne avant que ces contenus ne disparaissent. Une page Web hors d'usage courant est une page hors-ligne ou tout simplement disparue. Les archives du Web collectent des documents encore en usage courant pour éviter leur perte. Il s'agit d'une des premières différences avec les archives « classiques ». Bruno Bachimont propose d'utiliser le terme « collections » du Web pour éviter cet abus de langage. Comme les archives du Web conservent la production d'idées autour et suite à un événement, cet ensemble relève de la collection de bibliothèques plus que des archives. De leur côté, les archives tendent à conserver le lien entre le document et l'activité qui l'a produit afin de remonter jusqu'à la causalité de cette activité[119].

Puisque les archives du Web ne sont pas des archives à proprement parler, on comprend mieux pourquoi cette mission a été confiée aux bibliothécaires de la BnF plutôt qu'aux Archives nationales. De plus, l'encadrement légal des archives du Web a nommé la BnF et l'Ina comme dépositaires du dépôt légal du Web. L'intégration de ces archives dans une extension du dépôt légal implique un choix dans l'approche de ces futures collections, qui, « *avec l'évolution de la société, portaient une part de plus en plus importante de la connaissance et des échanges sociaux produits dans le pays*[120] ». Il ne faut cependant pas exclure l'influence de la réactivité des bibliothécaires face à la conservation du Web. En effet, ce sont deux bibliothèques nationales qui ont suivi l'exemple d'Internet Archive dès 1996, et depuis, ce sont toujours les bibliothèques qui prennent en charge l'archivage du Web. Sans doute y a-t-il eu une prise de conscience plus précoce de la part des bibliothécaires pour les nouveaux enjeux du numérique par rapport aux archivistes. Ces derniers étaient impliqués dans les réflexions autour de la numérisation de leurs fonds et toutes les problématiques qui en ont découlé, le Web n'était donc pas une priorité dans les évolutions des archives.

---

[118] CNRTL : https://www.cnrtl.fr/definition/archives

[119] MUSIANI Francesca, *et al.,* op. cit., p. 13.

[120] Annexe 3 : entretien à la BnF.

Malgré tout, il n'y a pas de rupture nette entre les archives physiques et les archives du Web. Toutes deux sont animées par bon nombre de problématiques communes. La question des doublons par exemple existe au sein des deux types d'archives. Ils ont d'ailleurs permis de conserver de nombreux documents à travers le temps et les destructions. Les collectes d'urgence ne sont pas propres aux archives du Web, elles sont pratiquées depuis longtemps par les archives physiques. Tout comme la problématique de la masse documentaire générée n'est pas propre au Web bien qu'elle y prenne une ampleur nouvelle. Les archives-collections du Web se sont simplement adaptées aux règles propres du numérique et du Web, à commencer par la notion de temporalité qui se dissipe avec l'absence d'un flux linéaire et continu. Cette instabilité du document numérique permet de le fractionner presqu'à volonté sans en perdre l'information ou la provenance puisque les métadonnées permettent d'assurer la compréhensibilité du document[121].

### 3.1.2. « Chaque archive Web est une reconstruction[122] »

La question sémantique n'est pas la seule différence entre les archives physiques et les archives du Web. Ces dernières ne peuvent être pensées comme des copies parfaites des pages qu'elles capturent. Une page Web n'est pas un élément homogène où l'ensemble du contenu est rattaché à un seul et même plan. Le corps de texte, les illustrations, les bannières, des *pop-ups*, les publicités, tout cela est rattaché à un code différent et tous sont reliés entre eux par un réseau de liens. Pour ces raisons, « il faut considérer la page moins comme une unité qu'un ensemble d'éléments, qui peuvent être collectés séparément[123] ». Lors de ses collectes, la BnF ne collecte pas l'intégralité des pages. Pour alléger ces fichiers, on tronque les pages en ordonnant au *crawler* d'ignorer les bannières, les illustrations, etc. Ces éléments sont rarement collectés plusieurs fois. On réutilise la plupart du temps une illustration collectée les années précédentes pour éviter les doublons et limiter le volume des collectes. D'autres éléments de page sont ignorés par Heritrix, car il n'a pas les capacités techniques pour les moissonner. C'est le cas des éléments codés en JavaScript par exemple, ou encore des polices et caractères qui peuvent être différentes une fois collectées, car le choix n'a pas été directement inscrit dans le code source du site Web[124].

---

[121] MUSIANI Francesca, *et al.,* op. cit., p. 28, 32.

[122] Ibid. p. 34.

[123] Ibid. p. 35.

[124] Ibid. p. 34-36, 43, 46.

De plus, une page Web n'a pas une temporalité unique et chronologique. Chaque élément peut être mis à jour indépendamment des autres. S'ajoute à cela la reconstruction des pages et liens dans les archives du Web[125]. Par exemple, lorsqu'on consulte la page d'accueil du *Monde* du 21 février 1999 dans la *Wayback Machine*, l'onglet « Nouvelles technologies » renvoie à la page du 8 février 1999. Cette dernière n'avait pas été collectée à nouveau. Le même type d'anachronisme se retrouve dans les illustrations. On peut citer la page d'accueil du site du CNRS qui arbore son logo endeuillé de noir sur une page d'août 2015 alors que ce bandeau noir n'a été ajouté qu'en novembre 2015 suite aux attentats de Paris[126]. Cela met en lumière le véritable patchwork que sont les archives du Web. Ces archives sont en quelque sorte falsifiées volontairement à cause des limites techniques, de stockage et les modes spécifiques d'évolution des documents sur le Web. En cela réside assurément la principale différence avec les archives physiques classiques. Là encore, Bruno Bachimont résume cette distinction en ces termes : « *Pour une archive traditionnelle, l'enjeu est de conserver un document comme produit d'une activité donnée, dont il est alors une trace probatoire, permettant de renseigner sur la nature de l'activité, de prouver les événements associés. Il est donc essentiel, pour entamer son exploitation de s'assurer que le document est bien le « bon », c'est-à-dire qu'il est bien ce qu'il prétend être : il doit être « authentique ». [...] L'authenticité repose sur l'intégrité. Pour une archive du Web, ce raisonnement ne peut plus tenir. En effet, l'archive du Web n'est pas le Web, l'archive d'un site n'est pas le site archivé. La raison essentielle tient à la nature même des contenus et des procédures de collecte : en particulier, la durée de captation étant supérieure au rythme de mise à jour du site, l'archive résultant de la collecte rassemble en fait des parties de site renvoyant à des temps ou époques différents du site : une partie correspond au site au temps $t^0$, une autre au temps $t^1$ après une mise à jour, etc. bref, le site archivé n'a jamais existé comme tel dans le Web*[127] ».

Puisque ces archives ne sont pas le reflet exact des sites originaux, comment signaler à l'usager ces particularités ? Plusieurs solutions sont à l'étude. Niels Brügger, par exemple, propose d'employer la philologie pour comparer les différentes versions d'une page Web. Une forme de diplomatique numérique se crée également. Ces *digital forensics* exploitent les données de navigation pour reconstituer le document critique et contextualiser la valeur de

[125] BRÜGGER Niels, « A brief outline of temporalities of the Web online and in Web archives », SCHAFER Valérie (dir.), *Temps et temporalités du Web*, Nanterre, Presses universitaires de Paris Nanterre, 2018, p. 65.

[126] MUSIANI Francesca, *et al.,* op. cit., p. 54-55.

[127] Ibid. p. 36.

chaque document numérique. Cela permettrait de combler les lacunes, dater les différents éléments, etc. pour signifier la part d'intégrité à l'usager[128].

## 3.2. Une valorisation complexe

Les collections formées grâce aux archives du Web font l'objet d'une réflexion autour de leur valorisation. Plusieurs éléments complexifient leur mise en valeur tout comme leur utilisation par des chercheurs ou tout usager.

### 3.2.1. L'indexation

Avant de penser à la valorisation des archives du Web, il a d'abord fallu convenir d'un mode d'indexation pour permettre d'identifier chaque élément et de pouvoir le retrouver par la suite. La masse considérable de données collectées et la complexité propre à l'instabilité du document numérique ont rapidement exclu l'idée d'une indexation manuelle. Les métadonnées permettent déjà de localiser et rassembler les éléments épars d'un document numérique. Il fallait donc un moyen pour identifier cet ensemble et le retrouver sur demande. L'indexation automatique a été retenue comme solution la plus adaptée et efficace. La BnF indexe donc automatiquement le fruit de ses collectes grâce à Heritrix, qui a été conçu comme un *crawler* et un robot d'indexation. L'indexation permet de retrouver les pages Web archivées depuis la plateforme de consultation des archives du Web[129].

Le premier mode d'entrée dans les archives de la BnF se fait par l'adresse URL du site. Cela a l'avantage d'éviter tout bruit documentaire et d'assurer que l'on obtienne directement la page Web recherchée. L'inconvénient majeur est qu'il faut connaître l'adresse URL exacte du site, ce qui n'est pas propice à l'exploration des archives ou à la recherche par analogie. L'utilisateur n'aura accès qu'aux pages et sites reliés à la page de départ par les liens hypertextes associés. La BnF n'est pas la seule à avoir privilégié, dans un premier temps, cette approche. Les archives du Web portugaises, britanniques ou japonaises appliquent aussi la méthode « *single URL approach* »[130].

Afin d'élargir les possibilités de recherche et d'exploitation des archives du Web, l'indexation plein-texte est souvent présentée comme une solution favorable. Le plein texte possède en effet plusieurs avantages. D'une part, il permet une utilisation plus intuitive du

[128] Ibid. p. 36-37.

[129] MUSSOU Claude, op. cit., p. 264.

[130] MUSIANI Francesca, *et al.,* op. cit., p. 58. ; GAME, Valérie, OURY, Clément, op. cit., p. 73.

portail de consultation, car il s'agit d'une indexation similaire à celle des moteurs de recherche. Mais contrairement à Google par exemple, les réponses à une requête sont ici neutres. Lorsque l'on fait une requête sur Google ou tout autre moteur de recherche à but commercial, les réponses apportées sont classées selon différents critères de fréquence de consultation, de sponsor, de lien renvoyant à tel autre site, etc. Dans une indexation plein texte pour les archives, les réponses sont classées selon le nombre d'occurrences des mots-clés de la requête et la position qu'occupent ces mots dans la structure du texte[131]. L'indexation plein texte est réalisée avec l'aide du logiciel NutchWAX, pour *Web Archive eXtension*, développé par l'IIPC[132].

L'indexation plein texte n'est pas pour autant dénuée d'inconvénient. Si cela permet une recherche plus large et intuitive à l'utilisateur, cela occasionne en parallèle un bruit documentaire non-négligeable. Même à l'échelle d'un corpus, les réponses sont à ce jour trop nombreuses et beaucoup d'entre elles n'ont pas d'intérêt vis-à-vis de la requête initiale. Une réflexion est en cours autour de la structuration de la recherche plein texte au sein de la BnF. D'autres institutions se sont penchées sur la question. L'Australie et son portail Australia Trove ont déployé l'indexation plein texte avec des résultats satisfaisants[133].

À ce jour, la BnF a déployé l'indexation plein texte dans quelques corpus. Les premiers tests ont eu lieu sur les collectes larges de 2006, 2007 et 2008 et les collectes ciblées autour des élections de ces mêmes années. Les collectes d'urgences des attentats de 2015 ont fait l'objet d'une indexation plein texte en 2016, motivé par la création du projet ASAP. La même année, les archives du Web des années 1990 ont bénéficié du même traitement[134].

### 3.2.2. Quels publics ?

Le service du dépôt légal numérique a travaillé assez tôt autour de la question du service aux utilisateurs de ses archives du Web. D'abord en créant un portail de consultation, puis en développant tout un ensemble d'outils d'analyse, d'extraction de données, etc. On ne peut pas sortir les données des archives du Web en dehors de l'enceinte de la BnF. Il s'agit de la contrepartie pour l'obligation aux éditeurs de sites de permettre la collecte de leurs données. Il est donc nécessaire de fournir aux utilisateurs de ces archives tous les outils dont ils peuvent avoir besoin pour mener à bien leurs recherches sur place[135].

---

[131] MUSSOU Claude, op. cit., p. 265.

[132] GAME, Valérie, OURY, Clément, op. cit., p. 74.

[133] ILLIEN Gildas, « Le dépôt légal de l'internet en pratique : les moissonneurs du Web », p. 23.

[134] MUSIANI Francesca, *et al.,* op. cit., p. 59.

[135] Ibid. p. 56.

La BnF a lancé plusieurs études et enquêtes auprès des utilisateurs des archives du Web et des chercheurs pour mieux cerner leurs besoins et utilisations. Une première enquête en 2006 avait été réalisée en partenariat avec Sciences Po pour observer la manière dont le panel d'étudiants, de chercheurs et de lecteurs s'approprie le corpus constitué autour des élections de 2002 et 2004[136]. L'étude de début 2011, menée par la délégation à la Stratégie et à la Recherche de la BnF, a dressé trois profils types parmi les utilisateurs : les chercheurs, principalement issus des sciences humaines, les professionnels comme les avocats, journalistes, documentalistes ou encore consultants marketing, et enfin le tout-venant de la bibliothèque de recherche. De son côté, l'Ina a aussi mené ses propres enquêtes et des ateliers pour les mêmes objectifs[137].

Le public des archives du Web de la BnF reste encore majoritairement composé de chercheurs. La recherche s'empare par ailleurs petit à petit des données mises à disposition. On peut citer l'exemple du parcours Web90 grâce auquel des chercheurs étudient le patrimoine numérique du Web des années 1990 en France ou encore des travaux un peu plus éloignés du numérique, comme une thèse menée sur les mémoires de l'immigration maghrébine en ligne[138]. Ce public vient consulter les archives du Web à la BnF pour leurs contenus, mais aussi pour la valeur ajoutée que fournit la BnF en proposant une aide à l'analyse qu'Internet Archive ne propose pas[139].

Enfin, une partie du public de ces archives n'est finalement pas encore née. Si elles servent à la recherche actuellement et à retrouver des informations disparues du Web, ces archives vont prendre toute leur mesure d'ici quelques décennies. Un des buts de l'archivage du Web est de conserver la production de documents numériques pour les générations à venir. Ce sera aussi à ce futur public de définir la portée et l'utilité de ces archives, qui s'adapteront alors à leurs besoins[140].

---

[136] GAME Valérie, ILLIEN Gildas, op. cit., p. 85.

[137] MUSIANI Francesca, *et al.,* op. cit., p. 69-71.

[138] Ibid. p. 76. ; JACQUOT Olivier, « Le web des années 1990 : 20 ans d'*Internet Archive* et 10 ans du dépôt légal du web en France », *Carnet de la recherche à la Bibliothèque nationale de France*, 19 novembre 2016. https://bnf.hypotheses.org/1309

[139] Annexe 3 : entretien à la BnF.

[140] Annexe 3 : entretien à la BnF.

### 3.2.3. Valoriser les collections

Le dépôt légal numérique diffère un peu dans le traitement de ses collections par rapport aux autres dépôts légaux. Lorsqu'ils reçoivent un nouveau document, ces services prennent en charge les entrées, le pointage et le catalogage. Les documents sont ensuite transmis aux différents départements de collections auxquels ils appartiennent. Ce sont ces départements qui assument la valorisation et la communication des documents. Dans le cas du dépôt légal numérique, la valorisation et la communication font aussi partie de leurs attributions. Ainsi, les postes de consultation présents au sein de la BnF et dans les BDLI sont à la charge du service du dépôt légal numérique qui doit veiller à leur bon fonctionnement et leur déploiement, épaulé par le DSI[141].

Pour valoriser les collections des archives du Web, des parcours guidés et des corpus ont été créés. Par exemple, les collectes ciblées autour des différentes élections du début des années 2000 ont fait l'objet d'un parcours qui a été mis en ligne en 2008, sous le nom de « Cliquez, votez : l'Internet électoral ». Ce parcours permet à l'usager de consulter plus facilement le corpus recueilli lors des collectes. L'entrée dans les collections est alors guidée, pour échapper à la recherche par URL seule. Ce parcours regroupe plusieurs thématiques comme « la caricature politique » ou « la stratégie de propagande des principaux candidats »[142].

Une valorisation sous forme d'événements ou de journées d'étude existe aussi. Plusieurs journées d'étude ont été réalisées au sein de la BnF, autour de certaines de ses collections. Ce fut le cas en novembre 2016 avec la journée d'étude intitulée « Il était une fois dans le web : 20 ans d'archives de l'internet en France », qui est revenue sur l'évolution des archives du Web. Des ateliers de découverte des collections ont également été organisés. Plus récemment, à l'occasion « Des Voix d'Orléans » en avril 2019, la BnF a animé un atelier de présentation et de découverte des archives du Web et du portail de consultation, à destination du grand public[143].

Cependant, la restriction d'accès, de consultation et d'exploitation des données des archives du Web est sans doute un frein à la valorisation de ces collections auprès d'un plus large public. De même, la recherche commence seulement à prendre en considération depuis quelques années la richesse documentaire qu'offrent les archives du Web[144].

---

[141] Ibid.

[142] ILLIEN Gildas, « Le dépôt légal de l'internet en pratique : les moissonneurs du Web », p. 26.

[143] Annexe 3 : entretien à la BnF

[144] Ibid.

## 3.3. Evolution des missions et des métiers

L'archivage du Web a nécessité quelques évolutions pour les missions de la BnF ainsi que pour les métiers des bibliothèques. Quelle a été l'ampleur de ces changements et sont-ils envisagés sur le long terme ?

### 3.3.1. Une spécialisation des métiers au sein du dépôt légal numérique ?

Au fil des expérimentations autour de l'archivage du Web à la BnF, la question de l'évolution des métiers des bibliothèques s'est posée. On a alors envisagé de créer de nouveaux postes, répondant aux nouvelles compétences numériques exigées. Une spécialisation des métiers au sein même du dépôt légal numérique a été envisagée et encouragée. Gildas Illien, ancien chef du service du dépôt légal numérique, a beaucoup écrit sur l'archivage du Web et les changements qui en ont découlé entre 2006 et 2011. Il a mis en lumière les nouveaux traitements documentaires que le volume exponentiel des collectes et des archives a rendu nécessaires. En effet, la sélection, la gestion et la valorisation de ces nouvelles collections impliquent une maîtrise technique de l'informatique ainsi qu'une connaissance approfondie des enjeux du numérique, en plus des compétences professionnelles des bibliothécaires. Les problématiques autour de la préservation pérenne du document numérique exigeaient d'être familiarisé aux différents formats en pleine conception, comme le ARC puis le WARC. De même, il fallait concevoir des solutions pour la consultation de ces archives tout en respectant les restrictions d'accès. Pour répondre à quelques-unes de ces évolutions, on a créé de nouvelles fonctions comme celle de chargé de collection numérique. Il s'agit d'un poste pour bibliothécaire ayant la responsabilité de rassembler les suggestions d'URL pour les collectes ciblées, envoyées par les bibliothécaires correspondants des départements de la BnF et des pôles associés. Ce poste demande aussi de planifier et de gérer au quotidien les collectes. Le poste d'opérateur numérique a également été créé. Il s'agit de gérer les processus automatisés des collectes et le traitement des collections. Il est aussi en charge de superviser l'indexation de masse faite par Heritrix et de gérer les formats et leur pérennité. En parallèle, l'expertise des bibliothécaires est aussi mise en avant. De par le volume de données archivées, les bibliothécaires du service du dépôt légal numérique doivent réévaluer les exigences documentaires pour correspondre au mieux à la représentativité des collections. Cela implique une légère concession sur les exigences scientifiques pour trouver un compromis entre la qualité des documents et la qualité de leur support de consultation. Les compétences documentaires des bibliothécaires sont aussi requises après la collecte pour définir, parmi la masse de données, ce qu'il faut valoriser en particulier pour leur qualité

documentaire et leur représentativité. Dans ces articles, Gildas Illien présente une division des missions de chargé de collection[145].

Près de 10 ans après les évolutions des métiers présentées par Gildas Illien et ses propositions pour l'avenir, la spécialisation au sein du service du dépôt légal numérique a-t-elle perduré ? Il s'avère que cette voie a été progressivement abandonnée. Suite à la suppression du département de la bibliothèque numérique en 2008, les compétences numériques se sont retrouvées réparties entre les différents départements de la BnF. Le rattachement du dépôt légal du service numérique au département du dépôt légal a été motivé par la mission commune et l'encadrement de la loi. L'idée d'une spécialisation du service du dépôt légal numérique a été abandonnée pour ne pas concentrer à nouveau les compétences numériques en un seul pôle et accorder un statut particulier aux collections numériques. Ainsi, l'essentiel des missions techniques du service est géré par le DSI. Ce sont aujourd'hui des bibliothécaires que l'on recrute pour le dépôt légal numérique, afin de bénéficier de leurs compétences documentaires, de gestion et de valorisation des collections numériques. La BnF est plutôt actuellement en voie de standardiser les profils recrutés pour ce service. L'attirance pour le numérique est favorisée, tout comme une expérience dans les collections numériques, mais ne sont pas un critère absolu[146].

Il faut relever qu'il est bien plus simple aujourd'hui de trouver un bibliothécaire à l'aise avec les questions du numérique et l'utilisation des outils adaptés qu'il y a encore 10 ans. La formation initiale des bibliothécaires et conservateurs y joue un rôle. Le numérique commence à s'insérer progressivement au sein des formations, que ce soit à l'ENSSIB ou à l'École des chartes. Une formation plus technique y est apportée, avec une meilleure appréhension des enjeux des collections numériques en bibliothèque. Cependant, il ne faut pas négliger les compétences personnelles de ces jeunes professionnels et des bibliothécaires plus généralement. Le numérique fait partie intégrante de notre société. Tout le monde, ou presque, est quotidiennement en contact avec l'informatique et le numérique. Chacun a développé ses propres compétences et connaissances tout en acquérant une plus grande aisance avec la diversité des supports et des outils. Cela est d'autant plus probant pour les plus jeunes générations[147].

---

[145] ILLIEN Gildas, « Le dépôt légal de l'internet en pratique : les moissonneurs du Web », p. 24, 26. ; GAME Valérie, ILLIEN Gildas, op. cit., p. 84-85.

[146] Annexe 3 : entretien à la BnF.

[147] Ibid.

### 3.3.2. Des compétences professionnelles encore difficiles à valoriser

Au sein de la BnF, onze personnes travaillent autour du dépôt légal numérique. Sept travaillent au sein même du service du dépôt légal numérique tandis que les quatre autres travaillent au DSI. Ces dernières travaillent presque exclusivement pour le dépôt légal numérique. En parallèle, le DSI accorde également deux temps partiels pour développer le dépôt légal des livres numériques. L'équipe du service du dépôt légal numérique est globalement restée stable dans le nombre de postes depuis la création du dépôt légal puisqu'en en 2008, sept personnes y travaillaient[148]. La proportion de titulaires et contractuels a varié au fil du temps, deux contractuels sont en poste actuellement. Si l'on compare avec les autres institutions de l'archivage du Web, la BnF est dans une situation favorable. En moyenne, une équipe pour l'archivage du Web se compose de 3,5 temps pleins et 5 temps partiels. En 2011, onze initiatives ne bénéficiaient que de temps partiels[149].

On constate un important *turn-over* au sein du service du dépôt légal numérique. Au moment de notre entretien, sur les sept membres, deux étaient en poste depuis plus de 10 ans. L'un d'eux a pris un nouveau poste peu de temps après l'entretien. Une autre personne travaille dans le service depuis 4 ans tandis que les quatre autres sont en poste depuis deux ans maximum. Deux recrutements devraient avoir lieu en septembre. S'il est tout à fait compréhensible de vouloir changer de service et de poste au bout de 10 ans, les mouvements plus récents sont intéressants. La principale raison identifiée pour ce *turn-over* dynamique réside dans les compétences techniques que l'on développe au sein du service. Ce sont principalement des compétences liées à de l'expertise et à la gestion de données numériques. Ces compétences sont encore difficiles à valoriser sur un CV dans le milieu des bibliothèques. Le lien direct avec le métier de bibliothécaire n'est pas évident à mettre en avant. On peut y voir la dernière forme de spécialisation, involontaire, au dépôt légal numérique[150].

On pourrait ajouter à cela la part inévitable d'idéalisme qu'il faut pour travailler pour l'archivage du Web. En effet, même après 23 ans d'Internet Archive et 13 ans d'archivage du Web par la BnF, l'intérêt de ces archives et collectes n'est pas totalement intégré et accepté par les acteurs du Web. Si dans l'ensemble, les collectes se mènent sans incident, la BnF reçoit malgré tout régulièrement des plaintes de la part d'éditeurs de sites, principalement lors des collectes larges. Ces éditeurs refusent que la BnF collecte et archive leurs données. Il s'agit souvent d'un malentendu, où la collecte est prise pour l'action d'un hacker. Une fois les raisons des collectes et l'enjeu de la conservation du patrimoine numérique expliqués, une partie

---

[148] ILLIEN Gildas, « Le dépôt légal de l'internet en pratique : les moissonneurs du Web », p. 24.

[149] GOMES Daniel, MIRANDA João, COSTA Miguel, op. cit., p. 413. ; Annexe 3 : entretien à la BnF.

[150] Annexe 3 : entretien à la BnF.

d'entre eux accepte volontiers de se soumettre à la loi, comprenant la mission de la BnF. D'autant que la restriction de la consultation de ces données leur garantit le respect du droit d'auteur. Une part de ces personnes cependant, reste opposée à cet archivage et à l'obligation que leur impose la loi DADVSI. Ils comptent souvent parmi les premiers acteurs du Web, avant le Web 2.0. Durant les années 1990 et le début des années 2000, les autorités souhaitaient exercer un contrôle strict sur le Web et ce qui y circulait. Cela s'est traduit par de la répression, des amendes, des peines de prison et du fichage auprès du ministère de l'Intérieur. Le souvenir du dépôt légal du livre qui recevait un exemplaire pour les fonds de la BnF et un pour le même ministère vient se mêler à la notion de dépôt légal numérique. Tous ces acteurs considèrent donc avec beaucoup de méfiance la mission de la BnF et y voient plutôt une entrave à l'idéal de la déclaration d'indépendance du cyberespace de 1996 qu'une véritable volonté de conserver la richesse de création de contenu du Web[151]. Travailler pour le service du dépôt légal numérique implique de devoir composer avec toutes ces réticences et de savoir que l'on ne sera pas toujours compris par les acteurs du Web. Cette reconnaissance parfois difficile peut être un frein supplémentaire à une carrière prolongée dans le service, s'ajoutant à la difficulté de faire valoir ses compétences techniques.

### 3.3.3. Quels liens avec les missions traditionnelles des métiers des bibliothèques ?

Les archives du Web sont le prolongement d'une mission traditionnelle de la BnF : le dépôt légal. Le service du dépôt légal numérique emploie de nouveaux outils et accorde une place importante et jamais égalée dans les bibliothèques à l'automatisation de la gestion des collections et des acquisitions. Malgré tout, c'est bien parce qu'il vise le même objectif de conservation du savoir et de la production d'idées qu'il a été rattaché au département du dépôt légal lors de la suppression du département de la bibliothèque numérique. La différence de traitement qui vise plutôt la représentativité que l'exhaustivité constitue un point de divergence, mais qui s'explique par l'étendu presque infini du Web et l'impossibilité de tout collecter. Les archives du Web constituent une collection numérique à part entière de la BnF et permet de faire évoluer en parallèle les métiers des bibliothèques en travaillant sur les nouveaux enjeux et outils du numérique et du Web.

Une grande partie du fonctionnement et des missions du dépôt légal numérique doit plutôt être perçue comme une évolution des missions traditionnelles que comme une rupture. Déjà en 2008, Gildas Illien avait cerné cette nuance en expliquant que « *pour les professionnels, il s'agit finalement de continuer à gérer un budget d'acquisition (qui s'exprime en octets et en nombre d'URL) dans un domaine thématique ou éditorial (qui échappe souvent*

[151] Ibid.

*aux classifications de Dewey) en utilisant des concepts ou des langages nouveaux (liés à la syntaxe des URL notamment)*[152] ». Le support de travail a changé et est en perpétuelle mutation, mais l'implication des bibliothécaires reste aussi essentielle. Cela passe par leur expertise documentaire et technique, mais aussi leur savoir-faire en terme de valorisation des collections et de service au public pour offrir la meilleure expérience de consultation aux différents publics.

[152] ILLIEN Gildas, « Le dépôt légal de l'internet en pratique : les moissonneurs du Web », p. 26.

# Conclusion

L'archivage du Web est encore très jeune. Cela fait à peine un peu plus de 20 ans qu'Internet Archive a commencé à archiver le Web mondial tandis que la Bibliothèque nationale de France archive le Web français depuis un peu plus de 10 ans. Cela fait peu de temps que ces archives ont acquis une stabilité dans leur encadrement juridique et leur gestion. L'archivage du Web évolue encore, notamment grâce au perfectionnement continuel de ses outils et formats de stockage.

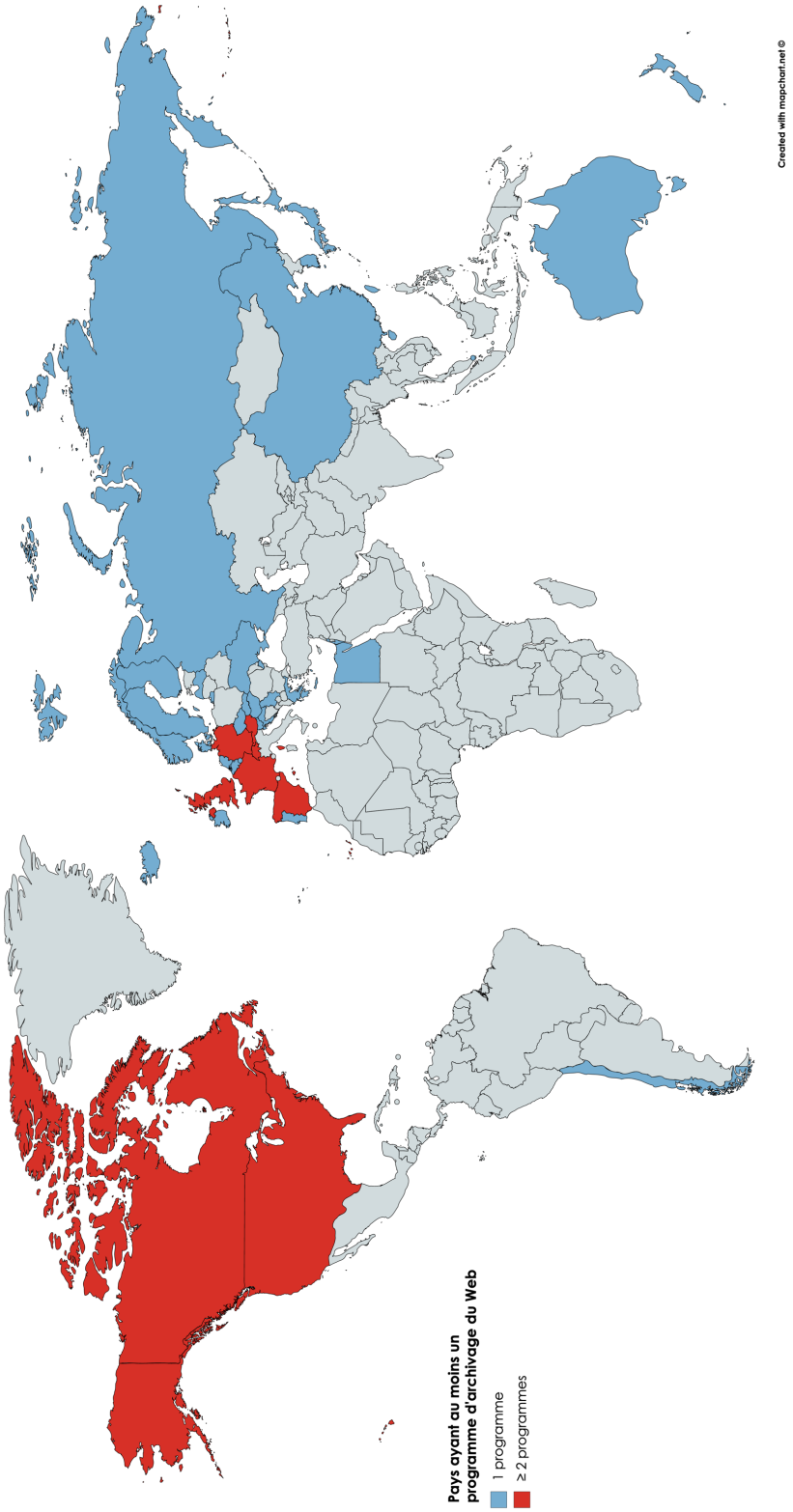
L'archivage du Web par la BnF n'a finalement pas tant fait évoluer les métiers des bibliothèques au sein du service du dépôt légal numérique. C'est principalement le support de travail qui a changé. Les méthodes ont dû être adaptées, mais au fond, le travail demeure essentiellement tourné vers l'acquisition, la conservation, la valorisation et le service aux publics. Le principal changement vient de la forme uniquement numérique de ces collections et à la nature éphémère du contenu disponible sur le Web. Il a fallu accepter que le dépôt numérique ne pourrait jamais être exhaustif. Archiver le Web, c'est comme essayer de retenir de l'eau avec les mains. On peut en contenir une partie, mais le plus gros s'écoulera inévitablement entre nos doigts. L'expertise professionnelle de longue date des bibliothécaires permet d'assurer que l'échantillon que l'on collecte soit représentatif de l'ensemble qui nous échappera toujours.

Un travail de reconnaissance de la valeur de cette expertise reste à mener au sein des métiers des bibliothèques. Une meilleure connaissance et sensibilisation autour de ce que sont les archives du Web de la BnF permettrait de valoriser les compétences plus approfondies que développées au sein du service du dépôt légal numérique. Cela pourrait effacer l'image d'un profil de bibliothécaire très spécialisé et cloisonné au numérique parmi les métiers des bibliothèques.

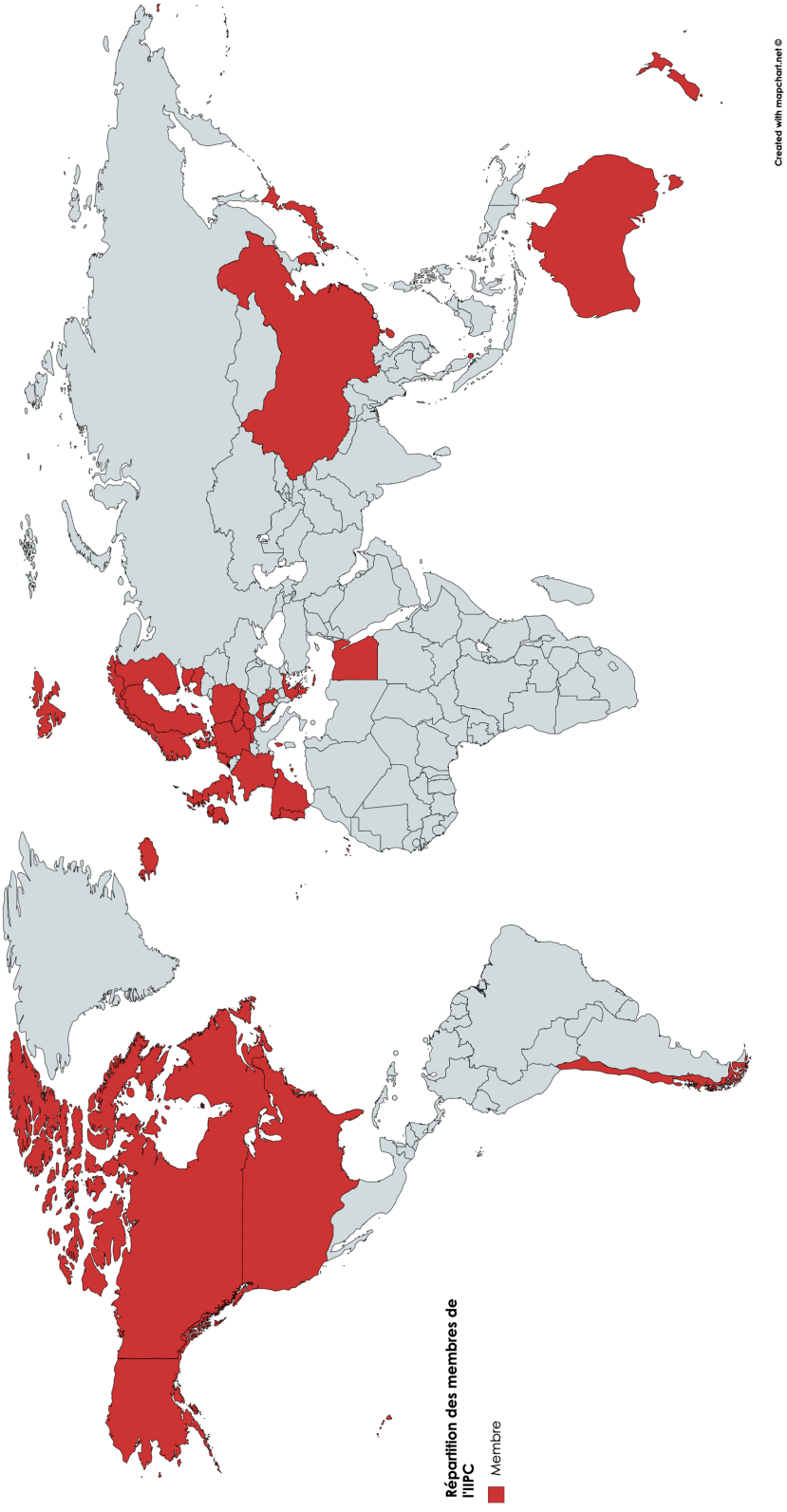
Pour finir, l'apport majeur de l'archivage du Web à la BnF réside certainement dans la mise en réseau des bibliothèques archivant le Web. Grâce à cette nouvelle mission, les bibliothécaires qui travaillent autour des archives du Web ont été amenés à développer une collaboration internationale durable. Ces bibliothécaires du monde entier travaillent main dans la main depuis plus de 15 ans pour créer les outils dont ils ont besoin, définir les meilleures stratégies de collecte et de conservation, partager leurs expériences sur l'encadrement juridique et la communication de leurs collections. Tout cela dans le but de partager leurs expériences et peut-être un jour partager leurs collections. On peut conclure en disant que la nouvelle Bibliothèque d'Alexandrie est finalement totalement numérique avec pour mission de conserver l'ensemble du Web mondial. Elle s'est décentralisée pour optimiser son rayonnement et assurer une conservation optimale de ses collections face aux différents risques matériels et physiques.

# Annexes

Annexe 1: Pays ayant au moins un programme d'archivage du Web.



Annexe 2 : Répartition des membres de l'IIPC.



Annexe 3 : Entretien avec le service du Dépôt légal numérique à la BnF, réalisé le 25 avril 2019.

AC : Le dépôt légal, loi de 2006 entrée en vigueur à la suite du décret de 2011. Des collectes BnF ont débuté, à titre expérimental, au tout début des années 2000. Ça s'est vraiment développé à partir des élections de 2002, avec un statut toujours expérimental. La BnF, donc, n'avait pas dans un premier temps les moyens techniques de réaliser ses propres collectes et nous avons pu faire appel à des prestataires extérieurs, notamment Internet Archive. La fondation Internet Archive qui archive le Web au niveau mondial depuis 1996. Nous avons fait appel à eux pour des collectes, et particulièrement des collectes larges, parce qu'en fait nos collectes, elles sont de deux types. D'abord la collecte large qui va chercher à moissonner tout l'Internet français. Alors pas avec un degré de complétude infaillible, mais pour des raisons techniques et pour des raisons liées à l'espace disque qui n'est pas illimité et également la bande passante et les ressources en terme de serveur qui ne sont pas illimitées non plus. Nous ne cherchons pas à collecter l'ensemble du Web français. Et nous nous limitons à un échantillon représentatif, ce qui est d'ailleurs inscrit dans la loi, contrairement au dépôt légal des livres, des périodiques, des documents audiovisuels sur support, des cartes et plans, etc. Le dépôt légal du Web français n'est pas exhaustif mais il est représentatif. C'est une première différence.

Ces collectes larges fonctionnent de la manière suivante. C'est-à-dire que nous récupérons ce qui ressemble à l'ensemble des adresses des sites français, l'ensemble des noms de domaine des sites français. Ce sont les .fr mais pas seulement, parce que ce sont aussi les .com, les .net, les .org, etc. basés en France. Le critère, ce n'est pas la langue, ce n'est pas le fait d'être en langue française, c'est le fait que l'éditeur du site soit basé en France et que le site... alors que le site soit hébergé en France ça c'est une autre... par exemple Ovh héberge plein de sites qui viennent, qui sont basés dans les pays de l'Est, etc. et qui n'ont rien à voir avec la France, c'est un hébergement purement technique, ça on ne les prend pas. Mais à partir du moment où l'éditeur ou le producteur du site est basé en France, on considère que nous prenons, même si la langue n'est pas du tout la langue française.

Et donc, les noms de domaine .fr mais également .com... Alors pour les .fr et puis également les noms de domaines liés au territoire d'outre-Mer, donc les .re, les .mq, etc., c'est un organisme, qui est l'AFNIC, et qui est chargé du registre des domaines liés à ces domaines nationaux. Donc .fr, .re, .mq, etc., là l'AFNIC nous procure la liste complète tous les ans. Et pour les .com, les .org, etc. c'est un peu plus compliqué, parce que nous sommes en relation avec différents bureaux d'enregistrements. Certains acceptent de nous donner leur liste, d'autres refusent. Et il n'est pas bien clair juridiquement que nous puissions les y contraindre. Donc on essaye de les convaincre mais ça ne marche pas à tous les coups et ça ne marche pas avec tout le monde.

On se sert aussi... quand on a d'autres moyens, que ce soit les moteurs de recherche, que ce soit des listes que l'on trouve ailleurs... de récupérer des noms de domaines français en .com, on le fait. Mais on n'a pas une exhaustivité totale à ce niveau là. Il nous manque notamment les .com français hébergés par Gandi qui est a priori le plus grand hébergeur français, donc on a quand même des manques assez significatifs. Donc la collecte large 2018, nous sommes partis sur une liste de 4 500 000 noms de domaine. Et nous estimons... le Web français entre 7 ou 8 millions de noms de domaine. C'est une pure estimation, c'est d'ailleurs assez difficile de trouver sur Internet ce genre d'estimation.

Océane Zielinski : Oui, j'ai commencé à chercher des chiffres, et oui c'est vrai que c'est... ce sont des estimations, ce n'est pas...

AC : On a un peu l'impression que ce sont des estimations à la louche et les sources se reprennent les unes les autres... Mais c'est assez difficile d'authentifier une estimation.

Nous estimons couvrir à peu près 60 % du Web français. Donc sachant que... nous avons en fait une taille maximale attribuée à chaque site, qu'on appelle le budget. Et ce budget il se traduit par un nombre d'URL. C'est-à-dire que là pour 2018, le robot de collecte avait pour consigne pour chaque nom de domaine, de s'arrêter lorsqu'il avait atteint pour ce domaine le nombre de 2 500 URL. Ce qui paraît beaucoup, et c'est effectivement beaucoup parce que ça permet de moissonner dans leur totalité à peu près 97 % des sites dont nous transmettons les noms au robot. Mais ça veut dire aussi qu'il y a 3 % des sites pour lesquels le contenu est plus important, et parfois beaucoup plus important que 2 500 URL. Si vous prenez par exemple la partie française de Wikipédia, c'est peut-être 4 millions et quelques d'URL. Donc nous n'en moissonnons vraiment que la partie la plus accessible à partir de la page d'accueil. La manière dont fonctionne le robot c'est effectivement qu'il simule le comportement d'un internaute humain. Quand il a le nom de domaine, il arrive sur la page d'accueil du site et il simule un clic sur tous les liens qu'il rencontre à partir de cette page. Et quand il y a une page qui s'affiche en réponse, pareil sur cette page il va balayer tous les liens possibles, jusqu'à ce qu'il est atteint le budget maximal de 2 500 URL.

OZ : Par exemple dans le cas de Wikipédia français, est-ce que d'une année sur l'autre, quand vous faites des captures, vous sélectionnez pour faire en sorte de prendre éventuellement les URL qui n'ont pas été sélectionnées les collectes d'avant ou c'est techniquement trop compliqué ?

AC : Alors on ne le fait pas pour les collectes larges. Par contre dans les collectes ciblées, ça on verra tout à l'heure, on a effectivement identifié des parties de Wikipédia qu'on cherche particulièrement à moissonner et qu'on va chercher à avoir dans nos archives. Mais dans la collecte large, on reste sur ce régime de représentativité en faisant un maximum de

chose de manière automatique. Et puis effectivement c'est un traitement de masse la collecte large : 4 500 000 URL... les humains interviennent mais pour de la surveillance, du contrôle, pour plus une maintenance technique. En fait l'aspect documentaire dans la collecte large, il est plus ou moins gommé, et volontairement en fait. Alors pour fixer les idées, la collecte large a occupé pour 2018 un volume un peu supérieur à 100 To. Voilà pour la collecte large.

Alors nous faisons aussi des collectes ciblées. Ces collectes ciblées, elles sont là pour le coup... elles sont le résultat d'une sélection sur des critères documentaires. Donc les différents départements de la BnF désignent en fait des personnes, des bibliothécaires qui ont pour mission de repérer des sites Web français particulièrement intéressants pour leur discipline ou leur domaine. Donc ça peut être des sites qui sont utilisés par leurs lecteurs, ça peut êtres des sites qu'ils repèrent en lisant la presse spécialisée de leur discipline. Après, nous le service du dépôt légal numérique, nous n'intervenons pas sur les critères documentaires. Nous considérons que c'est la politique documentaire de chaque département qui s'applique, et nous fixons juste des axes généraux... qui spécifient par exemple le caractère français des sites. Il faut quand même qu'ils soient dans le ressort du dépôt légal. Et puis le fait de ne pas prendre des contenus manifestement illégaux, qui contreviennent au droit d'auteur, etc., qui appellent à la haine raciale, à la violence, etc... Et donc ces sites, là pour le coup, ils sont le résultat d'une sélection humaine par des bibliothécaires. Les bibliothécaires de la BnF mais aussi les bibliothécaires d'institutions partenaires, notamment les bibliothèques du dépôt légal imprimeur, donc les BDLI, qui en région, comme la Bibliothèque d'Angers, comme la BM d'Angers... Alors toutes ne font pas des collectes, quelques fois elles participent aux collectes électorales pour leur région mais il n'y a pas véritablement de collecte régionale. Les BDLI donnent également accès à distance aux archives du Web avec les mêmes possibilités que si vous étiez dans les emprises de la BnF. Il y a également certaines BDLI qui font des collectes régionales, c'est le cas de l'Alsace, de la Loraine... je crois... Toulouse aussi, ils font une collecte. L'idée étant de sélectionner des sites pour les collectes ciblées qui sont à ce moment là moissonnés avec une profondeur plus importante. On peut avoir des limites de 45 000 ou de 200 000 URL. Et avec une fréquence aussi parfois plus importante, donc on a des collectes qui sont annuelles... alors la large est annuelle aussi... mais dans les collectes ciblées on a aussi des fréquence semestrielles, trimestrielles, mensuelles. Et en ce qui concerne les collectes liées à l'actualité hebdomadaire, quotidienne et il y a même une collecte qui se déroule plusieurs fois par jour, c'est le cas pour les réseaux sociaux. Particulièrement pour Twitter... pour suivre un événement qui donne lieu à beaucoup de tweets, comme à chaque fois on moissonne le contenu d'un écran, donc une vingtaine de tweets... donc là c'est utile de faire deux ou trois passages quotidiens parce que ça permet à ce moment là de voir...

OZ : En novembre 2015 je crois que ça avait été quatre collectes, quatre prises de vue par jour pour les tweets.

AL : Oui oui pour les attentats effectivement on avait adopté une fréquence plus importante. Les collectes ciblées, elles représentent entre 30 et 40 To selon les années. Et en nombre d'URL, c'est une trentaine de milliers d'URL sélectionnées par les bibliothécaires pour les collectes ciblées. À comparer aux 4 500 000 domaines de la collecte large. Donc c'est une partie du Web que l'on identifie comme étant... comme présentant un intérêt documentaire plus affirmé et qu'on choisit de collecter avec une profondeur et une fréquence plus importante. Mais là encore on n'est pas, on n'atteint pas ou très exceptionnellement l'exhaustivité. Dans cette collecte ciblée, il y a des sites de presse, il y a aussi des sites qui s'apparentent à de l'édition numérique. Des plateformes comme Cairn, comme Open Edition, comme Persée. Et là, je dirais que... et puis on a aussi des blogs, des carnets de recherches comme ceux de la plateforme Hypothèses. Et là en fait, c'est souvent le résultat d'une sélection, la plupart du temps quand on sélectionne la plateforme dans sa globalité avec nos outils, on a que des résultats qui sont très superficiels en terme de complétude. En fait on a ce qui est sous les feux de l'actualité, on a les revues ou les articles qui sont directement accessibles à partir de la page d'accueil et des pages directement liées à la page d'accueil. Mais dès qu'on veut rentrer un peu en profondeur, on atteint tout de suite les limites du budget même s'il est très élevé.

En plus, donc difficulté supplémentaire, souvent sur ces plateformes tout n'est pas accessible par des liens. La plupart du temps les contenus sont référencés dans une base de données dans laquelle il faut que l'utilisateur tape des mots clés, et il a en réponse les documents qui correspondent à sa recherche. Ça le robot ne sait pas le faire. Il ne va pas taper tous les mots-clés imaginables en fonction du domaine, de la discipline à laquelle se rattache le site. Et donc il y a un certain nombre, une part importante du contenu qui nous échappe à cause de ça. Se pose la question d'ailleurs de peut-être, dans des cas comme ça de développer la procédure du dépôt. Solliciter l'éditeur pour qu'il fasse un dépôt volontaire de son contenu. Donc ça on ne le fait pas à l'heure actuelle, mais ça fait partie des choses que l'on envisage, des évolutions que l'on envisage. Parce qu'effectivement on est bien conscient que nos technologies de collecte, pour des masses importantes de contenus comme celle-là, et bien ça ne permet pas effectivement l'exhaustivité et même dans les cas où l'on souhaiterait que ça soit exhaustif. Autre difficulté, les sites qui sont accessibles sur authentification. La loi nous permet de demander au producteur du site de nous créer un compte pour les besoins de dépôt légal et de nous fournir un login et un mot de passe pour qu'on puisse se connecter. Le robot de collecte Heritrix a des fonctions d'authentification automatique. Mais dans les faits c'est plus compliqué que cela parce que très souvent il y a des authentifications très complexes avec plusieurs critères, avec une session qui est créée, un identifiant qui est attribué et que le navigateur, quand c'est un navigateur, reprend automatiquement à chaque nouvelle requête. Le robot ne sait pas, ou pas dans tous les cas, faire ça. Et il y aussi un certain nombre de protocoles assez évolués de sécurité, que le robot n'est pas forcément capable de prendre en charge. En fait, quand il y a authentification, à quelques exceptions près comme les sites de

presse, on a ouvert un chantier particulier là-dessus, de fait bien que la loi nous y autorise, on renonce à collecter. Et là, en fait, ce sont les difficultés techniques qui font qu'on n'a pas vraiment la possibilité de nous connecter, parce qu'il faudrait pratiquement pour chaque site mener une instruction qui prendrait plusieurs jours, potentiellement plusieurs semaines. Dans la collecte large on ne collecte pas la partie protégée par un mot de passe lorsque les sites sont payants.

Voilà pour nos deux types de collecte. Pour revenir un peu sur l'historique, la BnF a mené des collectes expérimentales jusqu'en 2005-2006. Il y a eu une période entre 2006 et 2009 où la BnF faisait elle-même, avait développé les outils pour mener elle-même ses collectes ciblées, mais n'avait pas encore l'infrastructure nécessaire pour la collecte large. Donc là, la collecte large, sur ces années là, a été sous-traitée à Internet Archive. Et à partir de 2010 la BnF a mis en place l'infrastructure qui lui permettait d'être totalement autonome sur toutes les collectes, ciblées et larges. Dans nos collections, on a également... et ça c'est le blog ASAP dont vous parliez. Nous avons souhaité également avoir dans nos collections la partie du Web français qui correspondait à la période où la BnF n'archivait pas du tout le Web, même à titre expérimental. Donc on a acheté à Internet Archive la partie française de ses collections pour les années 1996 à 2000. Moyennant quoi nous avons des collections entre 1996 et aujourd'hui, même si la BnF n'a commencé à collecter que plus tard. On n'a pas, quand on regarde nos collections, la mention de l'origine mais effectivement tout n'a pas été collecté par la BnF. Il y a de l'Internet Archive et il y a aussi de l'Internet Archive pour le compte de la BnF, par un contrat, je sais pas si c'était vraiment une sous-traitance mais c'était une prestation qu'on commandait à Internet Archive.

Nos collections atteignent un volume global un peu supérieur à 1 pétaoctet. On a atteint ce pétaoctet à la fin de la collecte large 2018. Ça s'échelonne sur un peu plus de 20 ans, entre 1996 et 2019. Sachant quand même que la volumétrie est beaucoup plus importante pour les années à partir de 2010. *Grosso modo* une collecte large autour de 2005 c'était peut-être 10 To, c'était une trentaine ou une quarantaine en 2010 et c'est une centaine actuellement. Moyennant quoi le contenu récent est quand même beaucoup plus volumineux que le contenu ancien dans nos archives. C'est assez difficile à dire, mais on a quand même... quelques... et notamment dans le cadre de la journée d'étude de 2016, on a mené des travaux d'analyse de nos collections qui permettent d'avoir une idée de la répartition chronologique dans nos collections. Mais effectivement, ce qui est ancien est beaucoup plus... est présent en volume beaucoup plus réduit. D'ailleurs quand vous voyez les premières années, très souvent vous avez la page d'accueil et à un niveau de la page d'accueil et ça s'arrête là.

OZ : Oui, après on a augmenté le nombre de clic pour chaque site. Et pour les analyses que vous avez fait pour les journées d'étude de 2016, est-ce qu'il y a eu des publications qui ont été faites autour ?

AC : Vous avez le blog ASAP et vous avez aussi, mais ça c'est plutôt après, le blog WebCorpora. Ce que vous avez aussi, c'est une bibliographie des archives du Web qui était présente dans l'ancien site BnF, qui pour l'instant n'a pas été reprise dans le nouveau.

OZ : Je l'avais téléchargé.

AC : De toute façon vous pouvez l'avoir sur Internet Archive ou si vous allez à la BDLI d'Angers dans les archives du Web BnF mais je ne devrais pas vous le dire mais Internet Archive est plus simple parce que ça vous permet de télécharger.

OZ : Je me souviens que j'avais téléchargé le PDF justement de la liste, les 10 pages de bibliographie.

AC : Vous avez bien fait, vous avez bien fait ! C'est le réflexe, l'archivage personnel du Web reste quelque chose d'indispensable.

OZ : Là depuis que le site de la BnF a changé, il y a certaines pages où je me dis que j'aurais dû faire des captures d'écran. Tout n'est pas encore revenu dessus.

AC : Oui oui, mais là vous avez Internet Archive et les archives du Web BnF qui peuvent vous permettre... Là, la collecte est pas mauvaise. On n'est pas tout à fait exhaustif. Notamment les pages qui ont plusieurs onglets, quelques fois vous n'avez que le premier onglet qui est collecté, mais... Et puis les pages où vous avez des revues, avec une liseuse Flash ou JavaScript, là, la plupart du temps ça ne passe pas. Mais les pages qui sont en HTML CSS en général il n'y a pas de problème. Et les PDF non plus. Donc la bibliographie n'est pas perdue, grâce aux archives du Web et à Internet Archive !

OZ : J'avais pensé à aller consulter les archives du Web pour voir l'ancien site de la BnF pour certaines infos, le temps qu'elles réapparaissent.

AC : Il ne faut pas hésiter, c'est la preuve de l'utilité d'ailleurs de ce que l'on fait... Un autre exemple qu'on ressort assez souvent, c'est le site de l'Élysée. Quand le président change et que vous êtes un étudiant qui a le malheur de travailler sur la rhétorique dans les discours du président précédent... et bien très rapidement vous n'avez plus de quoi travailler parce que... Mais bon, là les archives du Web permettent dans certains cas de pallier ces manques. Mais on a trop souvent l'impression que ce qui est accessible sous forme numérique et Internet le sera pour l'éternité. Alors qu'on a très vite fait de débrancher un site ou de rendre du contenu inaccessible. On s'aperçoit qu'on a avec le numérique des outils qui sont encore beaucoup moins performants que pour le papier en terme de pérennisation.

26:56 [...] 27:14

OZ : Justement je me demandais, en terme de pérennisation des archives, si vous aviez réussi à trouver des solutions sur le court-moyen terme ? Pour le long terme je pense qu'avec le numérique c'est quand même un peu... compliqué.

AC : Oui ! Alors la BnF est, je dirais, en pointe sur le sujet puisqu'elle a mis en place un système qui s'appelle Spar, donc Système de préservation et d'archivage réparti. Ce système est un ensemble de serveurs, de baies de disques durs et de logiciels qui assurent la préservation des données numériques de la BnF. Alors données numériques de la BnF, c'est le résultat de la numérisation. C'est ce que l'on retrouve dans Gallica et Gallica intramuros et c'est aussi le résultat de l'archivage du Web. Spar a été mis en place en 2009. Nous versons systématiquement dans Spar toutes les données issues des collectes larges et ciblées. Donc ça c'est fait au fil de l'eau. Et on reprend progressivement les données qui ont été accumulées depuis les premières expérimentations d'archivage du Web de la BnF au tout début des années 2000 jusqu'à la mise en œuvre de Spar en 2010. La partie 2000-2009 est progressivement reversée dans Spar. L'intérêt de Spar c'est d'abord qu'il y a une garantie physique sur les fichiers. On n'est pas tributaire d'un support qui va se détériorer ou qui tout d'un coup deviendra illisible parce que les protocoles de lecture ont changé, etc. Là c'est vraiment sur des serveurs et les données sont répliquées à plusieurs endroits, donc en cas de *crash disk* on peut toujours récupérer les données. Les données sont aussi documentées en terme de format. Lorsqu'un fichier est intégré dans Spar, il fait objet d'une analyse du format. Chaque format est évalué en terme de pérennité. Il y a des formats qui sont refusés par Spar parce qu'ils ne présentent pas les garanties de pérennité et il y a aussi des fichiers sur lesquels on a des alertes parce qu'ils ne sont pas complétement conformes aux spécifications du format. C'est la raison pour laquelle les versements dans Spar ne sont pas des opérations neutres. Il y a un travail commun des informaticiens et des services producteurs de données pour statuer sur toutes les erreurs qu'on peut trouver. Quelques fois la solution est une conversion de format ou dans certains cas on stocke quand même tout en ayant conscience que le format n'est pas complétement pérenne. On a aussi des cas où on conserve les deux parce que la conversion va provoquer quand même, au moins à la marge, des pertes de données ou des modifications de données. Donc on conserve quand même le format non pérenne pour avoir les données dans leur totalité. A charge pour ceux qui voudront les consulter dans quelques dizaines d'années de se débrouiller avec les problèmes de conversion.

OZ : Oui et à eux aussi de refaire la veille pour permettre que ça dure un peu plus longtemps.

AC : Oui, oui et effectivement dans Spar, ça implique aussi une veille sur les formats y compris les formats d'il y a dix ans. Et une veille sur les outils qui permettent de visualiser ou de convertir ses formats dans d'autres formats.

Donc Spar c'est ça. C'est la sécurité physique des données et c'est un certain nombre de garanties que ces données vont rester lisibles et exploitables avec le temps. C'est aussi un certain nombre de métadonnées qui vont permettre à la fois de retrouver les fichiers et également de faire des statistiques sur nos données. De mieux connaître nos données. Et sur le site Web vous allez trouver des pages consacrées à Spar. Je pense que là aussi un petit tour dans les versions archivées va vous permettre d'avoir un contenu plus important.

Spar, c'est la mise en œuvre du protocole OAIS qui est un protocole international pour la préservation des données numériques. Donc voilà pour la préservation, mais on voit d'être tout à fait tranquille, puisque nos données historiques ne sont pas encore toutes dans Spar, et pour des raisons liées aux ressources humaines côté DSI, ça ne va pas forcément se faire tout de suite mais au moins c'est dans les tuyaux, petit à petit. Dans les années qui viennent on devrait pouvoir reprendre l'ensemble de nos données depuis les premières expérimentations jusqu'aux collectes actuelles.

OZ : Du coup, la partie qui avait été récupérée d'Internet Archive jusqu'en 2000 a aussi été basculée sur Spar ou pas encore ?

AC : Alors, on a basculé sur Spar... oui non là, en 2000... Alors oui ce qui est... Non là ça n'a pas encore été basculé. Ce sont des données qui sont en sécurité puisqu'il y a une version de toute façon de diffusion à laquelle vous pouvez accéder lorsque vous utilisez les archives du Web. Et là pour le coup c'est doublement en sécurité parce que ce sont des données qui viennent d'Internet Archive et dans l'hypothèse où nous, on perdrait ces données, on pourrait toujours les récupérer chez Internet Archive. Mais on va essayer quand même de préserver notre copie à nous (rire). C'est quand même ça le but du jeu.

OZ : Du côté de la durée de vie des serveurs, parce qu'on l'estime à plus ou moins 10 ans on va dire, est-ce...

AC : Oui, je pense qu'à la BnF les serveurs sont changés plus souvent que ça. Plus proche de 5 ans que de 10 ans. Parce qu'en fait on attend pas qu'ils soient à toute extrémité, enfin il y a des cas exceptionnels, où il y a ça dans beaucoup d'organisations, une appli développée pour des besoins très spécifiques qui n'a pas évoluée, qui ne tourne qu'avec des systèmes d'exploitations très anciens, donc on laisse sur une veille machine ou on virtualise et on met l'ensemble sur une machine moderne mais dans un environnement qui simule un environnement beaucoup plus ancien en priant pour qu'à chaque modification de configuration, ça veuille quand même bien redémarrer. Mais les serveurs, on a plutôt tendance à les changer

plus vite parce qu'on est juste en ressources et les applications deviennent un peu lentes bien avant que le serveur soit physiquement usé. Donc là les serveurs sont changés tous les 5 ans, en gros. C'est difficile à dire, là c'est vraiment la main du DSI, on n'a pas vraiment la main là-dessus. Mais 10 ans ça me paraît effectivement beaucoup, sauf exception comme celle dont je vous parlais tout à l'heure. Et encore, dans ce cas là, le serveur physique a quand même changé... On a en général de meilleurs résultats en utilisant un serveur récent avec une machine virtuelle qu'en gardant un serveur ancien.

OZ : Il y a combien de personnes qui travaillent dans le service du dépôt légal numérique.

AC : Dans le service du dépôt légal numérique, on est 7. Et au DSI on a 4 personnes qui sont, on va dire, plus ou moins dédiées aux activités du dépôt légal numérique. Alors ces 4 personnes c'est côté dépôt légal du Web. Sur une autre de nos missions qui est la mise en œuvre pour l'instant à titre expérimentale du dépôt légal des livres numériques, là on a 2 autres personnes au DSI mais ce n'est pas deux équivalent temps plein, c'est 2 personnes qui consacrent une part de leur temps à ces chantiers là.

On est dans une situation, au regard des ressources humaines, qui est plutôt favorable par rapport aux institutions qui font la même chose que nous à l'étranger. On est plutôt dans ceux qui ont du personnel. Malgré ça, ça ne suffit pas pour faire tout ce qu'on veut et notamment à suivre toutes les évolutions du Web. Voir notamment comment évoluent toutes les technologies de structuration des pages, d'accès aux données, les évolutions de JavaScript, les nouvelles bibliothèques comme Node.js, Symphony, etc. C'est très très intéressant d'un point de vue informatique et d'un point de vue ergonomique : le Web est beaucoup mieux qu'avant. Mais il devient aussi beaucoup plus difficile à collecter avec nos robots, les technologies que nous avons. Là il y a un chantier, je dirais, de veille d'évolution des outils qui est très important et qu'on a pas forcément les moyens de mener parallèlement à nos missions de base. Même si, en se comparant à nos homologues étrangers, on n'a pas du tout à nous plaindre des moyens dont nous disposons.

OZ : J'aurai plus une série de questions qui portent sur le côté missions de la bibliothèques et métiers. Dans quelle mesure les missions du service du dépôt légal numérique et Web correspondent aux missions globales de la BnF et en quoi elles en diffèrent aussi un peu éventuellement ?

AC : Le dépôt légal... alors c'est pas pour rien qu'effectivement le service dépôt légal du numérique appartient au département du dépôt légal. Ça n'a pas toujours été le cas. Lorsque ce service a vu le jour début des années 2000, il a d'abord été rattaché à un département qui à l'époque s'appelait le département de la bibliothèque numérique. Et ce département il a

cessé d'exister en 2008 parce que la direction de l'établissement à constaté qu'avec le numérique, avec l'évolution de la technologie était partout et était dans tous les départements. Donc ça n'avait pas de sens, ça n'en avait plus de faire un département de la bibliothèque numérique. On considérait que ce département en ayant fait essaimé le numérique un peu partout dans l'établissement, en ayant contribué à cette diffusion, il avait terminé sa mission et que donc le numérique devait prendre place dans l'ensemble des départements de l'établissement en fonction de ce à quoi il était utilisé. Moyennant quoi Gallica s'est retrouvé au département de la coopération, puisque Gallica n'est pas seulement la bibliothèque numérique de la BnF, elle est aussi celle de nombreux partenaires à Paris et en régions.

Les archives du Web ont été rattachées au dépôt légal parce qu'on a considéré que le moissonnage du Web était un mode d'entrée pour des collections numériques, que la loi nous donnait mission de collecter. Parce que ces collections, avec l'évolution de la société, portaient une part de plus en plus importante de la connaissance et des échanges sociaux produits dans le pays. C'est vrai qu'à partir de peut-être 2005, le nombre d'internaute en France a été suffisamment important pour qu'on ne puisse plus considérer qu'en gros tout ce qui était sur Internet existait aussi sur le papier. On a eu des documents qui ont eu une existence uniquement numérique et qui auraient été perdu à tout jamais si on ne collectait pas.

Donc les archives du Web se situent dans le droit fil d'une mission traditionnelle de la BnF, qui est celle du dépôt légal. Avec quand même une particularité, parmi les services du département du dépôt légal, le service DLN est le seul qui s'occupe également de l'accès à ses collections. Parce que les livres ou les périodiques dans le département nous nous occupons des entrées, du pointage, du catalogage, de la veille sur ce qui n'a éventuellement pas été déposé, mais une fois que ça a été enregistré, catalogué, et bien on transmet les documents aux différents départements de collection qui auront pour mission de les valoriser et de les communiquer. Et ça n'est plus notre affaire. Alors que pour le Web, dans tous les départements et dans les BDLI, il y a des postes de consultation des archives du Web, mais ces postes de consultation c'est nous, DLN et DSI, qui sommes responsables de leur bon fonctionnement et de la manière également dont les lecteurs de l'établissement vont accéder au contenu. Donc là il a une petite spécificité, c'est qu'effectivement nous sommes le seul service du DDL à s'occuper aussi de l'accès à nos collections.

L'autre point de divergence, l'autre particularité, c'est cette question de la représentativité et non pas de l'exhaustivité. La volumétrie des contenus disponibles sur le Web et un certain nombre d'obstacles techniques à une collecte complétement exhaustive font qu'on a pris ce parti de la représentativité plutôt que l'exhaustivité. Et puis il y a un certain nombre de chose du point de vue du signalement aussi. Les documents qui arrivent par dépôt légal, que ce soit des livres, des périodiques, des estampes, des documents sonores sur supports, et bien ils sont catalogués et ils sont signalés au catalogue, en général à la pièce. Bon, on a des cas particuliers de documents traités en recueil comme les manuels scolaires. Mais le cas général c'est quand même des documents traités à la pièce. Ce n'est pas le cas de

ce que l'on moissonne au titre du dépôt légal du Web, puisque si vous voulez consulter un document que nous avons moissonné, vous devez connaître l'URL ou vous devez passer par l'un des parcours guidés avec effectivement... mais ça couvre une infime partie des collections que nous avons moissonné.

OZ : Je n'ai pas encore eu le temps de regarder vraiment dans le détail mais j'avais vu que, justement, pour les collectes de 2015 la recherche plein texte qui a été expérimentée.

AC : Il y a effectivement une recherche plein texte sur pour l'instant trois corpus, donc les attentats 2015, le Web des années 90 et actualité 2010-2018. Et on va d'ici la fin de l'année travailler sur un autre corpus, le corpus élections. Du moins on va traiter le début du corpus, les années 2002-2007. Sachant que si l'on va de 2002 à 2007 on doit avoir quelque chose comme 30 To de données. Et ça, avec la recherche plein texte, ça pose toutes sortes de questions, notamment celle du bruit. Si vous faites quelques tests déjà sur un corpus plus limité comme attentats 2015, vous avez assez facilement des réponses que vous n'attendiez pas, des réponses qui n'ont pas à être là, etc. il y a tout un travail sur la structuration des index, sur la sélection des données indexées qui est assez critique pour que les données soient véritablement utilisables ensuite et répondent vraiment aux besoins des utilisateurs. La recherche plein texte ce n'est pas si simple que ça à mettre en œuvre. Alors après, d'autres bibliothèques à l'étranger, on observe ça avec attention, ont mis ça en place et notamment de manière récente l'Australie. Qui a quand même des volumes importants, 600 To, ils archivent depuis très longtemps, ils ont commencé en 1996. Je n'ai pas eu le temps de regarder précisément, mais c'est apparemment très intéressant ce qu'ils ont fait. En plus ils ont une interface, une espèce de portail qui s'appelle *Australia Trove* qui est un mélange de Gallica, du CCFR et des archives du Web. Mais vous pouvez très facilement isolé un corpus web. Je pense qu'il y a vraiment beaucoup de choses intéressantes à observer et éventuellement à mettre en œuvre dans ce qu'ils ont fait.

OZ : Pour la valorisation des archives du Web auprès du grand public, même si c'est une part infime, est-ce qu'il y a des projets envisagés pour essayer de familiariser les gens avec la plateforme qui n'est pas forcément très intuitive ou pratique.

AC : Vis-à-vis du grand public, c'est une difficulté. Tout le monde peut y accéder. Tous les lecteurs ayant une carte recherche peuvent y accéder mais les stats de consultations montrent que les lecteurs sont encore assez peu nombreux à y accéder. On a, selon les mois, entre 1 000 et 2 000 consultations, en cumulant les consultations BnF et les consultations de l'ensemble des BDLI, donc ça reste marginal, très marginal par rapport à l'ensemble de nos collections. Encore faut-il préciser que dans ces consultations, il y a la presse et parmi les chantiers du dépôt légal du Web, on a depuis 2014 un chantier qui est « presse payante », qui

consiste à récupérer au titre du dépôt légal les versions PDF des différentes éditions locales des titres de la presse régionale. Ouest-France par exemple, vous avez une cinquantaine d'éditions locales. Maintenant on ne les reçoit plus toutes sur papier. Celle que l'on reçoit sur papier au titre du dépôt légal c'est l'édition de Rennes et uniquement celle-là. Les autres, donc si vous voulez lire l'édition de Concarneau ou l'édition de Morlaix, il faut passer par les archives du Web et aller consulter le PDF de l'édition qui vous intéresse. C'est tout le bénéfice parce que la presse occupe une place démentielle dans les magasins physiques et pour les journaux, les éditeurs de journaux, c'est une contrainte assez pénible que de devoir acheminer quotidiennement toutes les éditions. Donc c'est intéressant pour eux et c'est intéressant pour nous de substituer à ce dépôt légal papier un dépôt légal numérique. Moyennant quoi les lecteurs qui vont consulter ces éditions locales consultent les archives du Web même s'ils n'étaient pas du tout partis de chez eux en se disant « tiens je vais consulter les archives du Web à la BnF ». Ils voulaient consulter un titre de presse et on les oriente vers les archives du Web. Donc effectivement, on a encore une fréquentation qui est circonscrite à des usages très exceptionnels. Soit des chercheurs identifiés, soit quelques lecteurs qui utilisent de leur propre initiative, mais ça reste marginale. Il faut préciser quand même qu'Internet Archive répond aux besoins de beaucoup d'internautes qui ont... depuis 2010 sur le Web français on moissonne plus de choses qu'eux, mais eux moissonnent quand même pas mal de choses. Donc dans beaucoup de cas je pense qu'Internet Archive répond aux demandes des internautes qui ne viennent pas chercher ici. Et effectivement on est plus identifié par des labos de recherche qui viennent chercher l'accès aux collections mais aussi, souvent, ce qui les intéresse ce sont les traitements informatiques que l'on peut faire sur ces collections, les extractions et les traitements préparatoires aux opérations de fouilles de données que leurs équipes vont mener sur nos données. C'est ça aussi qui les intéresse, pas seulement nos collections. Ce sont nos collections, avec une valeur ajoutée matérialisée par certains traitements.

OZ : C'est ça aussi qui distingue un peu plus les archives du Web des archives papier, plus classiques puisqu'il a vraiment la dimension de l'exploitation des données numériques avec tout un outil qu'il n'y a pas pour le papier.

AC : Oui, même si la consultation du papier est de plus en plus mêlée au numérique, avec des outils comme Gallica, avec le catalogue et aussi data.bnf où vous avez la possibilité de récupérer des ensembles de données. Même quand les documents que vous consultez sont sur papier, vous avez de plus en plus de traitement numérique à effectuer lorsque vous faites des recherches. C'est une particularité parce que nous sommes accessibles que comme cela mais c'est général.

OZ : Pour les bibliothèques des pôles associés, avez-vous des formations pour les bibliothécaires en charge de la présentation et l'explication des archives du Web ?

AC : On a commencé en 2014 le déploiement. Il y a 26 bibliothèques qui sont pôles associées, dépôt légal imprimeur. Actuellement on est à 18 bibliothèques déployées. A chaque fois qu'on déploie les archives du Web dans une bibliothèque, on fait une formation pour les bibliothécaires de la BDLI qui seront en charge de la collecte et de l'accompagnement du public pour la consultation des archives du Web. Donc c'est systématique, il y a une demie journée de formation que l'on assure lorsqu'on déploie une BDLI. On est aussi là pour répondre à leurs questions quand il y a besoin ou pour organiser des événements de valorisation. Début avril à Orléans, par exemple, il y a eu quelque chose. L'une des collègues du service s'est déplacée pour faire une présentation dans le cadre d'un événement plus global. Je crois qu'il y a eu une présentation en commun avec l'Ina, l'Ina étant en charge du dépôt légal de tout ce qui est Web TV et Web radio. Donc on fait aussi des événements de valorisation autour de nos collections, y compris dans les BDLI.

OZ : Autour de la valorisation des collections, qu'est-ce qui reste le plus compliqué, quelle(s) difficulté(s) techniques qui a éventuellement été surmontée mais qui a posé problème ?

AC : Comme outil de valorisation, on a un outil assez basique qui est un ensemble de parcours guidés, thématiques. La valorisation, ce sont aussi des événements ou journées d'études, participation à des colloques, ça peut être des publications. On est aussi un service qui intéresse un peu de temps en temps les journalistes, en particulier quand il y a des questions liées à des données qui disparaissent sur le Web, on a quand même de temps en temps des sollicitations pour interviews... Ce qui est peut-être quand même un frein à la valorisation, c'est le caractère assez restreint de la diffusion de nos collections puisque les lecteurs doivent être physiquement ici ou dans une BDLI pour pouvoir consulter nos collections. Et qu'ils n'ont pas la possibilité d'imprimer ou d'extraire du contenu. Ça découle de la loi sur le dépôt légal qui nous donne cette mission de moissonner le Web, qui s'oppose à ce que les producteurs de site le refusent mais en contrepartie, la consultation de ces collections se fait uniquement sur place et dans les BDLI pour des lecteurs, pour des chercheurs accrédités. C'est ce qui nous permet de travailler mais c'est aussi ce qui rend difficile la consultation de nos collections. Même si le réseau de BDLI commence à être assez étendu. Mis à part en Bretagne où il faut aller à Rennes. Dans le sud-ouest on a aussi... alors Bordeaux a été déployé, mais pour X raison, notamment technique et l'impossibilité pour l'instant de trouver un terrain d'entente avec le service informatique de la ville, ça a été déployé mais ça ne marche pas. Donc il faut aller à Toulouse ou à Poitiers. Quand on est à Bordeaux, ce n'est quand même pas tout près. L'obstacle essentiel est là quand même, c'est-à-dire ce caractère

un peu difficile d'accès et difficile d'utilisation de nos collections. Un obstacle secondaire étant l'absence d'une recherche plein texte sur l'ensemble du corpus. Même si, je pense que cette recherche plein texte, si elle existait, serait aussi génératrice de toute sorte de déception mais on voit que là où elle existe, par exemple en Australie, ça ouvre quand même à beaucoup d'autres usages des archives du Web. Eux, en plus, ont mis en ligne pour tout internaute. Donc d'ici on peut consulter les archives du Web australienne. Ça aussi ça permet d'avoir une audience beaucoup plus importante. Je ne dis pas qu'on ne fera jamais ça ici, mais il faut une sacrée évolution de la loi ou un risque politique et juridique clairement assumé au plus haut niveau dans l'établissement pour que ça se fasse. Alors après on peut aussi dire que le public de nos collections du dépôt légal, ce sont des gens qui ne sont pas encore nés. Parce que quand on regarde les collections papier, on communique beaucoup de documents datant du XIX$^{e}$ siècle ou d'avant. Donc on va avoir des gens en 2050, 2100, 2200 qui vont peut-être s'intéresser de très près à ce à quoi ressemblait le Web français au tout début du XXI$^{e}$ siècle. Donc notre public c'est celui-là aussi, le public de l'avenir. La raison d'être du dépôt légal, d'une manière générale, c'est de donner aux chercheurs de demain les matériaux de leur recherche. Mais bon, ça n'empêche pas de vouloir que nos collections soient le plus vite possible utilisées, trouvent le plus vite possible leur public.

OZ : Du point de vue des profils qui travaillent dans ce service, est-ce qu'au moment de l'embauche, ce sont des profils vraiment très spécifiques qui sont recherchés, avec des compétences techniques déjà acquises ? Ou est-ce que la formation se fait plus ou moins sur le tas ?

AC : Alors en fait, on a, je ne parle pas de l'équipe DSI, je parle uniquement des personnes qui sont au service DLN. Ce sont des profils de bibliothécaires, même si on reste un service où il y a actuellement 2 contractuels dans l'équipe, il y en a eu beaucoup plus que ça avant, précisément parce que c'était un profil assez peu commun en bibliothèque, qui n'attirait pas ou qui faisait peur aux bibliothécaires lambda. Sachant que maintenant, que ce soit l'ENSSIB ou l'École des chartes, les formations sont quand même devenues beaucoup plus techniques, et je vois que les bibliothécaires des générations plus jeunes ont beaucoup moins d'appréhension par rapport à toutes ces questions. Donc jusqu'à présent les recrutements se sont fait... Effectivement on a privilégié les candidats qui, par leur expérience, pouvaient avoir eu affaire à des collections numériques, mais ça n'est pas une obligation. On a aussi des gens qui étaient dans d'autres départements de la BnF et qui sont ensuite venus ici. Donc c'est un service qui a eu une très très forte spécificité, y compris du point de vue du recrutement. Mais on est plutôt en voie de standardisation et les recrutements sont de plus en plus ouverts à des gens qui sont des bibliothécaires avec une attirance particulière pour le numérique et un petit peu quand même d'expérience autour des collections numériques. Mais on va chercher de

moins en moins de profils purement techniques. Alors pour le DSI ça reste évidemment un impératif.

OZ : Les formations ENSSIB et École des chartes préparent-elles mieux maintenant à ces spécificités des nouveaux métiers des bibliothèques, avec la part grandissante du numérique, ou est-ce qu'il y a encore du chemin à faire ?

AC : Il y a peut-être encore du chemin à faire, mais on constate quand même que le numérique prend une importance croissante dans les programmes de ces écoles. Alors c'est peut-être aussi parce que les élèves eux-mêmes ont une meilleure connaissance de ces technologies. On n'observe pas non plus d'année en année l'évolution des programmes, on a plutôt tendance à voir qui en sort. Et effectivement, il y a sans doute beaucoup moins d'appréhension par rapport au numérique maintenant qu'il y a encore 10-15 ans.

OZ : Selon vous, est-ce qu'une spécialisation au sein des métiers des bibliothèques pour les parties qui traitent plutôt des collections numériques serait souhaitable ou est-ce que garder un aspect plus polyvalent reste plus intéressant sur le long terme ?

AC : Alors oui, la spécialisation c'est ce à quoi on a renoncé en supprimant le département de la bibliothèque numérique en 2008. L'idée c'était de dire le numérique est partout et que tout le monde doit faire du numérique et que ce sont des compétences qui ont tendance à se banaliser dans le bon sens du terme et qu'il ne faudrait surtout pas recréer une espèce de chapelle numérique dans un établissement qui est tout papier et qui veut le rester. La spécialisation, quelques fois est nécessaire, mais l'idée est plutôt de faire admettre que les collections numériques sont les collections de l'établissement, au même titre que les collections papier. Qu'il doit y avoir... ça les lecteurs sont aussi assez réceptifs à ça et souvent ce qui les intéresse c'est d'accéder à un certain contenu. Certains vont vouloir absolument que ce soit sur papier, mais la plupart du temps s'ils étaient venus avec l'idée de consulter un article de périodique sur papier et qu'on leur dit « ah oui mais dans telle base de données vous allez avoir cet article », ils sont très contents de le lire sur écran ou de l'imprimer, parce que c'est quand même parfois plus agréable de lire sur papier.

OZ : Au sein de l'équipe du département, est-ce qu'il y a un roulement dynamique pour les renouvellements de postes ou est-ce que le service a plutôt acquis une sorte de stabilité dans la durée des postes ?

AC : Il y a quand même... ce n'est pas un service où on reste 20 ans. Bon on n'existe pas depuis 20 ans d'un autre côté (rires). Mais... on a parmi nos collègues 2 personnes qui sont là depuis plus de 10 ans mais dont l'une va s'en aller la semaine prochaine prendre un

autre poste. Sinon on a une personne qui est là depuis 4 ans et puis les autres sont arrivés dans les deux dernières années. C'est un service avec un *turn-over* assez important... Là on va avoir effectivement assez vite deux postes non pourvus qui vont être mis au mouvement, avec peut-être des recrutements en septembre pour les deux postes en janvier. Et donc des gens qui seront encore plus nouveaux que les nouveaux qui sont là depuis 2 ans et qui vont devenir les anciens. Donc oui, on va dire qu'il y a une rotation assez rapide. D'où l'idée qu'il faut effectivement pouvoir intégrer rapidement de nouvelles personnes, de nouveaux profils, et pas forcément des profils très techniques. Il faut aussi que des personnes qui sont dans d'autres départements de la BnF et qui ont une attirance pour le numérique puissent venir et être accueillies dans le service.

OZ : Est-ce qu'il y a une ou des raisons qui ont été identifiées pour expliquer ce *turn-over* ?

AC : Alors... peut-être que... les compétences que l'on développe dans ce service, qui sont plus des compétences liées à de l'expertise, sont un petit peu difficiles à valoriser ensuite sur le marché du travail des bibliothèques. Le danger, si on reste trop longtemps, c'est de se retrouver un peu catalogué comme une espèce d'expert ne sachant faire que ça. Certains collègues m'ont fait part de ce genre de difficultés. Et puis quelqu'un qui est là depuis 10 ans, c'est tout à fait légitime qu'il ait envie de faire autre chose. On est dans un établissement où il y a une telle diversité d'activités que l'on peut changer radicalement de métier simplement en changeant de tour. Les collègues ont bien raison de profiter de ces possibilités. Mais effectivement... le numérique reste, de fait, une spécialisation. Bon je dirais que la BnF est aussi un peu spécialiste de l'appel à des compétences très pointues sur des micro-spécialités, que ça soit les métadonnées... C'est très intéressant d'avoir occupé quelques postes comme cela dans une carrière, mais les gens n'ont pas forcément envie de faire ça toute leur vie. On peut tout à fait comprendre...

OZ : Quelle est la part d'idéalisme dans l'archivage du Web et tout ce qui se fait autour ?

AC : Ah ! Si vous voulez, il y a une espèce de mythologie de l'internet qui a commencé à la fin du XX$^{e}$ siècle. Vous avez en 1996 la déclaration d'indépendance du cyberespace. Et le fait de rendre... faire de tout ça un objet de patrimoine est un enjeu passionnant. Même si ceux qui sont peut-être les idéologues du Web ont peut-être une certaine hostilité à l'intervention de... de l'État, de la puissance publique là-dedans et voient peut-être notre activité un peu comme du flicage ou de l'atteinte à la vie privée. Il y a quelque chose d'assez ambigu là-dedans. Parmi les acteurs du Web, on est perçu de manière assez diverse, assez

contrastée. Certains sont convaincus de l'utilité de ce qu'on fait, mais d'autres, d'autres se méfient et nous rangent plutôt du côté du Ministère de l'intérieur.

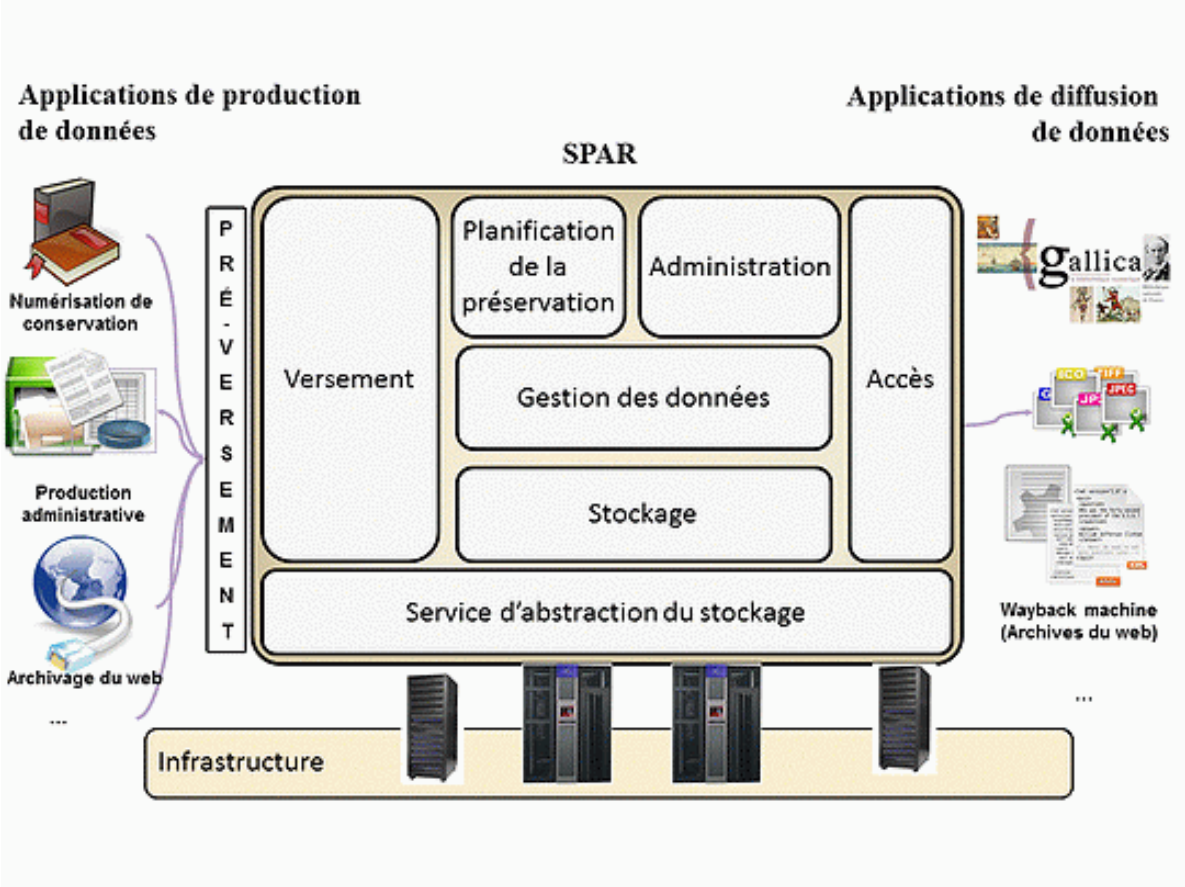
Alors il faut un peu d'idéalisme pour se lever le matin en disant « je vais archiver le Web français pour le siècle des siècles », mais il faut savoir qu'on ne sera pas toujours bien compris, bien perçus... par les acteurs du système. D'ailleurs dans nos opérations de collecte, ça c'est plutôt valable pour la collecte large, on a assez souvent des plaintes de producteurs de sites qui nous prennent d'abord ou pour des hackers ou pour des voleurs de données qui veulent siphonner le contenu de leurs serveurs. Ils disent qu'ils ne sont pas du tout d'accord pour qu'on prenne leurs données. Alors certain, mais pas tous, disent « ah oui bien sûr allez-y » quand on leur explique que c'est pour des besoins de conservation et pas de flicage... mais ça ne va pas de soi en fait. Et d'ailleurs le dépôt légal, maintenant... je parle du dépôt légal en général, y compris celui des livres... c'est uniquement à des fins de conservation mais encore il y peut-être 20 ans, il y avait un exemplaire transmis au Ministère de l'intérieur. Donc il y avait quand même une dimension de censure là-dedans. Et dans le cas particulier de l'Internet, il y a quand même un historique assez lourd, où en particulier au début de l'Internet, fin des années 90 début des années 2000 où la puissance publique s'intéressait surtout à Internet pour... à des fins de répression ou pour coller des amendes ou pour mettre en prison des hébergeurs qui avaient eu le malheur d'avoir sur leur serveurs des photos de telle ou telle star, etc. On a eu des hébergeurs qui se sont retrouvés avec des centaines de milliers de francs, à l'époque, d'amendes qu'ils ne pouvaient absolument pas payer. Et certains des acteurs de l'Internet actuel, qui étaient déjà des acteurs de l'Internet de cette époque là, n'ont pas oublié. Donc nous on récupère un peu tout ça.

OZ : Finalement, est-ce que l'archivage du Web serait une nouvelle forme de bibliothèque d'Alexandrie, un peu comme celle qui est recréée en Égypte ?

AC : Oui, alors on espère que les serveurs ne vont pas brûler (rires). Mais en principe il y a quand même du stockage dans des endroits différents. Mais oui, peut-être., en tout cas c'est une composante de... Alors est-ce que la BnF est une nouvelle bibliothèque d'Alexandrie... peut-être. On a quand même maintenant plus de mal à avoir... à rassembler dans une seule collection quelque chose qui ressemblerait à toute la connaissance du monde. Je dirais que maintenant c'est plutôt en réseau. Parce que la Bibliothèque nationale de France, comme son nom l'indique, elle est nationale et limitée à la France. Même si dans nos collections il y a pas mal d'ouvrages étrangers reçus par acquisition... si on veut... si on veut de l'allemand, il vaut mieux aller à la bibliothèque nationale et universitaire de Strasbourg qu'à la BnF. Si on veut de l'anglais il vaut mieux aller à la British Library ou la Library of Congress. C'est quand même une partie de la connaissance. On est plutôt maintenant dans un fonctionnement en réseau et on a des outils qui permettent de localiser les documents dans d'autres bibliothèques, mais je dirais que la nouvelle bibliothèque d'Alexandrie est répartie

entre un certain nombre de grands établissements dans le monde... et que l'Internet contribue largement à orienter les lecteurs vers le... le bon établissement. On a des catalogues collectifs qui permettent ça mais... Mais oui je vois plutôt ça comme ça. Les bibliothèques, maintenant, c'est plutôt le travail en réseau qu'un établissement qui aurait tout.

Annexe 4 : SPAR, schéma fonctionnel. **©BnF**



# Table des matières

# Table des illustrations

# Table des annexes

RÉSUMÉ

Depuis 1996, une partie du Web mondial est archivée par différentes institutions. La Bibliothèque nationale de France archive le Web français depuis 2011, à travers plusieurs types de collectes. Depuis quelques années, on note un intérêt nouveau de la recherche pour ces archives, à travers leur fonctionnement, leur conservation et leur exploitation par les chercheurs. Cependant, aucune étude récente n'a été menée sur l'influence qu'a pu avoir cette nouvelle mission de la BnF sur les métiers des bibliothèques.

Ce mémoire propose de retracer les grandes lignes du développement de l'archivage du Web pour ensuite aborder plus en profondeur les procédures d'archivage et de conservation utilisées par la BnF. Cette étude se fera à travers l'analyse des deux collectes d'urgence réalisées autour des attentats de Paris en 2015 et proposera également une actualisation de l'influence de cette mission d'archivage sur les métiers de bibliothèques au sein de la BnF.

**mots-clés :** archivage du Web, archives de Web, Web, Internet, Bibliothèques nationale de France, métiers des bibliothèques, conservation numérique, patrimoine numérique

ABSTRACT

Since 1996, a part of the global Web is archived by various institutions. The French Web is archived since 2011 by the French National Library, through different collecting ways. These archives have recently drawn the attention of research, interested in how they work, how they are preserved and used by research itself. However, no study has been led yet to measure the impact of the archive management led by the French National Library on the different missions and jobs of libraries in general.

This essay is trying to tell the history of the development of Web archives and leads to the analysis of archive procedures and conservation employed by the French National Library. This study will first zoom in on the two emergency collects carried out in 2015 after the Paris Attacks, and will also suggest an update on how the archiving mission impacted the work achieved in the core of the French National Library.

**keywords :** Web archiving, Web archives, Web, Internet, French National Library, library professions, digital conservation, digital heritage

Présidence de l'université
40 rue de rennes – BP 73532
49035 Angers cedex
Tél. 02 41 96 23 23 | Fax 02 41 96 23

# ENGAGEMENT DE NON PLAGIAT

Je, soussigné(e) ZIELINSKI Océane
déclare être pleinement conscient(e) que le plagiat de documents ou d'une partie d'un document publiée sur toutes formes de support, y compris l'internet, constitue une violation des droits d'auteur ainsi qu'une fraude caractérisée.
En conséquence, je m'engage à citer toutes les sources que j'ai utilisées pour écrire ce rapport ou mémoire.

signé par l'étudiant(e) le **06** / **06** / **2019**

**Cet engagement de non plagiat doit être signé et joint**
**à tous les rapports, dossiers, mémoires.**

Présidence de l'université
40 rue de rennes – BP 73532
49035 Angers cedex

université angers