Archivage du Web. Suivi du projet d'archivage des sites Internet des campagnes électorales de 2012 au titre du dépôt légal

Anece Oubaidourahaman

Université Paris-Est Créteil Val de Marne

UFR de Lettres et Sciences humaines

Département d'histoire

Archivage du Web

Suivi du projet d'archivage des sites Internet des campagnes électorales de 2012, au titre du dépôt légal

Rapport de stage pour l'obtention du Master 2 Professionnel « Histoire et médias »

Anece Oubaidourahaman

Promotion 2011-2012

Tuteur universitaire : Madame Claire Blandin, Maîtresse de conférences en histoire contemporaine

Maître de stage : Monsieur Clément Oury, Chef du service du Dépôt légal numérique

Stage : Bibliothèque nationale de France, service du Dépôt légal numérique (département du Dépôt légal)

Session de septembre 2012

Remerciements

Je tiens particulièrement à remercier Clément Oury, Sophie Derrot, Annick Lorthios, Géraldine Raison, Peter Stirling et tous les autres membres de l'équipe du dépôt légal Web, pour leur disponibilité, leurs conseils et leur soutien dans mon travail au quotidien.

Je remercie également tout le personnel du département pour m'avoir accueilli au sein de la BnF.

SOMMAIRE

LISTE DES SIGLES ET DES ABRÉVIATIONS

AFNIC	Association française pour le nommage Internet en coopération
BCWeb	BnF collecte du Web
BDLI	Bibliothèque de dépôt légal imprimeur
BnF	Bibliothèque nationale de France
DADVSI	(Loi relative) au droit d'auteur et aux droits voisins dans la société de l'information
DAV	Département audiovisuel (BnF)
DEP	Droit, économie, politique (département) (BnF)
DCO	Département des collections (BnF)
DDL	Département du dépôt légal (BnF)
DLN	Dépôt légal numérique (service) (BnF)
DLWeb	Dépôt légal du Web
DSI	Département des systèmes d'information (BnF)
IIPC	International Internet Preservation Consortium
INA	Institut national de l'audiovisuel
NAS	NetarchiveSuite
PHS	Philosophie, histoire, sciences de l'homme (BnF)
SED	Service Étude et développement (BnF)
SPAR	Système de préservation et d'archivage réparti
SSP	Service Support et production (BnF)
URL	Uniform resource locator

INTRODUCTION

Le service du Dépôt légal numérique au sein du département du Dépôt légal de la Bibliothèque nationale de France (BnF) m'a accueilli pour mon stage de fin d'étude du 13 février au 27 juillet 2012. Ce stage avait pour principale mission le suivi du projet d'archivage des sites Web traitant des campagnes électorales de 2012.

Ce projet s'inscrit dans une double continuité. D'une part, il s'intègre dans le processus de constitution de corpus aussi bien en science politique qu'en sciences sociales et en histoire, qui correspond à l'une des missions traditionnelles de la BnF. D'autre part, il représente une suite logique des opérations d'archivage effectuées à l'occasion des élections de 2002 (présidentielle et législatives), 2004 (régionales et européennes), 2007 (présidentielle et législatives), 2009 (européennes) et 2010 (régionales). Ces collectes ont été réalisées dans le cadre du dépôt légal de l'Internet, qui vise à constituer un échantillon représentatif des publications françaises en ligne.

En 2012, la Bibliothèque s'est de nouveau lancée dans une opération de capture à grande échelle des sites politiques. Cette collecte a concerné divers types de sites (sites officiels, sites de presse, blogs, réseaux sociaux...) et plusieurs types d'émetteurs (sites de partis, de candidats ou encore de militants). L'archivage des sites a été réalisé en lien avec de nombreux partenaires, notamment avec les bibliothèques de dépôt légal imprimeur. C'est dans ce cadre que le service du dépôt légal numérique chargé du pilotage technique et organisationnel du projet, a recruté un stagiaire.

L'objectif du projet est de conserver la trace de l'un des événements les plus actifs de la vie politique française. En effet, depuis 2002, le rôle joué par les « Web-campagnes » dans le déroulement d'une élection prend de l'ampleur à chaque nouveau scrutin. L'évolution la plus marquante depuis 2002 est sans doute l'apparition des réseaux sociaux (ni Facebook, ni Twitter n'existaient en 2002) qui font désormais parti intégrante des outils de campagne électorale.

Le Web a certainement facilité la production et la diffusion de contenu, notamment dans le cadre des élections où même un petit candidat indépendant peut au minimum être présent sur la Toile par l'intermédiaire d'une page Facebook. Cependant, comme le rappelle

Clément Oury[1], ces contenus mis en ligne souffrent d'une certaine volatilité. En effet, il suffit que l'auteur se désintéresse de son site, que la redevance annuelle pour le nom de domaine (c'est-à-dire l'adresse où se situe le site) ne soit pas acquittée, ou que l'hébergeur technique rencontre des difficultés, pour que les contenus disparaissent ou que des liens soient brisés. Cette particularité du Web concerne particulièrement les contenus publiés dans le cadre d'une campagne électorale. En effet, ces sites ont pour principale vocation de peser sur le scrutin, une fois l'échéance passée, ils n'ont donc plus de raison d'exister.

Ainsi les sites électoraux présentent la double particularité d'être très dynamiques durant un temps relativement court, celui de la campagne, et de connaître un fort risque de disparition dès la fin du premier tour d'un scrutin. Cette temporalité implique pour les organismes en charge de la collecte, la mise en place d'une structure de type projet : le projet Elections 2012. Il s'agit d'une collecte ciblée qui porte sur une sélection de sites repérés par des bibliothécaires, choisis en raison de leur thème ou de leur rapport à l'événement donné.

Pour archiver un site Web, il existe plusieurs possibilités comme la capture d'écran ou encore la capture vidéo mais la solution la plus efficace demeure le moissonnage. En effet, grâce à cette méthode, les fichiers constitutifs d'un site sont récoltés par l'intermédiaire de robots logiciel nommés *crawlers* (Heritrix[2] à la BnF). A partir d'une liste d'adresse URL[3], ce logiciel navigue sur les sites comme le ferait un internaute surfant sur tous les liens d'un site Web. Tous les contenus rencontrés par le logiciel sont copiés. Le paramétrage du logiciel et sa surveillance par des professionnels permettent de choisir la fréquence et la profondeur de capture pour chacun des sites. Les archives sont ensuite stockées sur disques ou sur bandes dans des entrepôts numériques (copies en plusieurs exemplaires, suivies de l'évolution des formats de fichiers, émulation…). L'accès aux collections s'effectue par l'intermédiaire de la Wayback Machine[4] qui s'apparente à un navigateur. Ce dernier permet de surfer dans les archives sur toutes les anciennes versions de site collectées à partir d'une URL donnée.

Le service du dépôt légal numérique assure le paramétrage et la surveillance des collectes. Ce service joue un rôle d'intermédiaire entre les informaticiens et les

[1] Oury, Clément, « Soixante millions de fichiers pour un scrutin, les collections de sites politiques à la BnF », *Revue de la Bibliothèque nationale de France*, Paris, n° 40, 2012

[2] Heritrix est un logiciel open source utilisé pour archiver le Web : http://crawler.archive.org

[3] *Uniform Ressource Locator* : désigne une chaîne de caractères utilisée pour adresser les ressources du Web

[4] Wayback Machine : développé par Internet Archive, cette application permet d'accéder aux données Web archivées. Plus d'informations sur la Wayback Machine : http://archive-access.sourceforge.net/projects/wayback/ , consulté le 7 septembre 2012.

bibliothécaires : d'une part, il travaille en collaboration avec des ingénieurs du département des Systèmes d'information, d'autre part il est associé à une équipe de bibliothécaires spécialisés, répartis dans les départements thématiques de la BnF, qui effectuent des sélections de sites en fonction de leurs disciplines.

Collecte, indexation, conservation et communication, l'archivage du Web permet d'aborder sous un angle différent tous les enjeux liés à l'univers des archives. D'autre part, le Web pose également de nouvelles questions. En effet, la Toile contient des données de natures très diverses : texte, image, vidéo, sons… Il est donc difficile de définir une unité documentaire. Il serait logique de définir le site Web comme l'unité de référence, toutefois ce dernier est assez difficile à appréhender d'un point de vue technique. Si on peut aisément comprendre ce qu'est le site d'une bibliothèque ou d'une université, comment peut-on considérer un blog hébergé sur une plateforme ? Est-ce un site à part entière ? Le site ne serait pas plutôt la plateforme d'hébergement ? Néanmoins, Niels Brügger parvient à distinguer « cinq strates analytiques différentes sur le Web[5] » : l'élément Web (la plus petite unité, correspond par exemple à une image), la page Web (ce qui est visible sur une fenêtre du navigateur), le site Web (constitué de plusieurs pages Web), la sphère Web (plusieurs sites abordant un même thème, on peut par exemple parler de Web campagne ou de Web politique) et enfin le Web pris dans son ensemble (normes, institutions…).

Ce stage permettait donc de confronter les connaissances acquises au cours du premier semestre au sein d'un domaine assez particulier, celui du Web : de quelle manière parvient-on à archiver les sites Web ? Comment documenter et conserver ce type de document ? Le stage étant centré sur la collecte des sites Web portant sur les campagnes électorales de 2012, nous aborderons les questions posées par ce biais.

Dans une première partie, nous présenterons le dépôt légal numérique au sein des collections du dépôt légal de la BnF.

Dans une deuxième partie, nous étudierons la collecte des sites Web traitant des campagnes électorales de 2012. En effet, la mission principale du stage étant le suivi du projet Élections 2012, il est nécessaire d'aborder les enjeux de cette collecte : pourquoi archiver ces sites ? Comment y parvenir ? Quel est le public visé ? Comment valoriser cette collection particulière ?

Enfin, la dernière partie est consacrée à la description des missions et des activités du stage.

---

[5] Brügger, Niels, « L'historiographie de sites Web : quelques enjeux fondamentaux », *Le Temps des médias*, 2012/1 n° 18, 2012, p. 159-169.

I. Le dépôt légal numérique à la Bibliothèque nationale de France : une collection dans la continuité des fonds

A. Présentation de l'organisation générale de la BnF

L'effectif de la bibliothèque est réparti entre trois grandes directions (la Direction des Collections, la Direction des Services et des réseaux et la Direction de l'Administration et du personnel) et cinq délégations (Délégation à la Diffusion culturelle, Délégation aux Relations internationales, Délégation à la Communication, Délégation à la Stratégie et à la recherche et Délégation au Mécénat).

1. Deux directions chargées des activités bibliothéconomiques

Les activités bibliothéconomiques se divisent principalement entre la Direction des Collections et la Direction des Services et réseaux.

La Direction des Collections (DCO) a la garde des collections, la mission de les traiter et de les mettre à la disposition du public. Ainsi la DCO est responsable des acquisitions (excluant le dépôt légal des livres et des périodiques) et de la gestion des salles de lecture thématiques. La DCO est constituée des départements thématiques et de services transverses. Lors du stage, j'ai pu visiter le département Droit, économie, politique et le département Philosophie, histoire, sciences de l'homme. Ces deux départements ont participé à la sélection des sites Web qui ont été collectés dans le cadre du projet Elections 2012.

Quant à la Direction des Services et réseaux (DSR), elle compte notamment le département du Dépôt légal, qui produit la Bibliographie nationale française, et elle a la responsabilité de la conservation, la coopération et de la reproduction. Par ailleurs, elle est également responsable de toute l'informatique de l'établissement. La DSR compte les départements suivants :

- département Information bibliographique et numérique
- département de la Conservation
- département de la Coopération
- département du Dépôt légal
- département de la Reproduction

- département des Systèmes d'information

Les départements de la Direction des Services et des réseaux et ceux de la Direction des Collections traitent, reçoivent, orientent, cataloguent, relient et restaurent les collections du dépôt légal imprimé (DSR), du dépôt légal sur autres supports (estampes, musique, cartes et plans, audiovisuels) et des acquisitions (DCO). Ces deux directions sont donc complémentaires et vouées à travailler ensemble. Si l'on prend pour exemple le dépôt légal du Web, les départements thématiques de la DCO sont chargés de sélectionner les sites selon leurs disciplines. Au sein de la DSR, le département du dépôt légal, associé au département des Systèmes d'information, est chargé de répondre à cette demande.

2. Le dépôt légal à la BnF

Le département du Dépôt légal (DDL) est responsable de la collecte ainsi que du traitement bibliographique des imprimés et des documents numériques en ligne. Les documents spécialisés sont quant à eux traités par les départements thématiques de la DCO (Audiovisuel, Cartes et plans, Estampes, Musique). Ainsi, le DDL est chargé de la gestion de la collecte, qui consiste à gérer les flux de documents, depuis la réception jusqu'au catalogage et à la bibliographie. Quant aux tâches de conservation, de valorisation et de communication des documents, elles relèvent des départements des collections. Le DDL est organisé en six services répartis en fonction du type de documents traités :

- Bibliographie nationale française / périodiques
- Bibliographie nationale française / livres
- Service du Dépôt légal numérique
- Service de la Gestion des périodiques
- Service de la Gestion des livres
- Service de Redistribution documentaire

Institué par François I^{er} en 1537, pour les imprimés, le dépôt légal a été progressivement étendu aux estampes, cartes et plans (1648), aux partitions musicales (1793), aux photographies (1925), aux phonogrammes (1938), aux affiches (1941), aux vidéogrammes

(1975), à la télévision, à la radio et aux logiciels (1992) puis enfin à l'Internet en 2006[6]. Le dépôt légal a pour objectif de rassembler la mémoire du patrimoine culturel diffusé sur le territoire français. Ainsi, toute publication publiée ou diffusée en France est assujettie au dépôt légal. Il s'agit donc d'une collection de référence sur la production éditoriale française.

Le dépôt légal est régi par le Code du patrimoine pour les aspects législatifs[7], et par le décret n° 93-1429 du 31 décembre 1993 et des arrêtés de 1995 et 2006 pour les aspects réglementaires. Le Code du patrimoine charge la BnF de collecter, conserver et communiquer[8] :

- « les documents imprimés, graphiques, photographiques, sonores, audiovisuels, multimédias, quel que soit leur procédé technique de production, d'édition ou de diffusion, [...], dès lors qu'ils sont mis à la disposition d'un public » ;
- « Les logiciels et les bases de données […] dès lors qu'ils sont mis à disposition d'un public par la diffusion d'un support matériel, quelle que soit la nature de ce support. » ;
- « […] les signes, signaux, écrits, images, sons ou messages de toute nature faisant l'objet d'une communication au public par voie électronique. »

Le dépôt légal est historiquement l'obligation pour chaque producteur de contenu culturel (imprimeur, éditeur) de déposer des exemplaires (le nombre variant selon l'importance du tirage) de ses travaux auprès de la bibliothèque royale, puis nationale. Avec l'arrivée de l'Internet, il était nécessaire d'adapter le dépôt légal à l'univers du numérique. La loi (votée dans le cadre de la loi DADVSI[9], en août 2006) et son décret d'application (paru en décembre 2011) ont donc été intégrés dans le Code du patrimoine, aux côtés des autres dispositions relatives au dépôt légal[10]. Le législateur a opté pour une définition assez large (« signes, signaux, écrits…par voie électronique ») volontairement sans évoquer les termes d'Internet ou de Web. En effet, le dépôt légal concerne tout ce qui circule en ligne et qui ne relève pas de la

[6] Les pages dédiées sur le site de la BnF apportent plus d'informations à ce sujet : http://www.bnf.fr/fr/professionnels/depot_legal.html, consulté le 14 septembre 2012.

[7] Code du Patrimoine, articles L131-1 à L133-1 et R131-1 à R133-1.

[8] Code du patrimoine, article L131-2, modifié par l'Ordonnance n°2009-901 du 24 juillet 2009 - art. 5.

[9] Loi n° 2006-961 du 1er août 2006 relative au droit d'auteur et aux droits voisins dans la société de l'information.

[10] Le décret d'application partage le dépôt légal de l'Internet entre la BnF et l'Institut national de l'audiovisuel (INA), voir I. B. 1.

correspondance privée. En ce sens, le dépôt légal du Web ne repose sur aucun jugement de qualité et se situe donc dans la continuité des collections sur papier.

B. Le dépôt légal du Web

1. Cadres, objectifs et moyens

Bien que le cadre légal ait été défini par la loi DADVSI, la BnF a mené différentes opérations d'archivage du Web dès 2002. Les collections des archives de l'Internet remontent même jusque 1996, car l'établissement les a acquises auprès d'Internet Archive[11].

La loi de 2006 complétée par son décret d'application partage le dépôt légal du Web entre la BnF et l'Institut national de l'audiovisuel (INA). En effet, l'INA a la charge de collecter les sites du domaine de la communication audiovisuelle parmi les sites français, il s'agit :

- des sites émanant des services des médias audiovisuels,
- des Web TV et Web radios,
- des sites principalement consacrés aux programmes radio et télévision,
- des sites des organismes de l'environnement professionnel et institutionnel du secteur de la communication audiovisuelle.

La BnF a quant à elle pour tâche de collecter tout le reste de l'Internet français.

Par ailleurs, le décret d'application de décembre 2011[12] définit également ce qu'est l'Internet français, un champ difficile à déterminer. Autant pour le document papier, il est facile de définir un périmètre de dépôt légal basé sur le territoire national[13], autant il est complexe de définir les frontières de la Toile puisque l'intégralité des contenus Web est diffusée sur le territoire français. Le décret a tout de même fixé un périmètre français. Ainsi, le dépôt légal du Web concerne :

[11] Internet Archive est une organisation américaine à but non lucratif consacrée à l'archivage du Web. Les collections de cette organisation sont accessibles en ligne : http://archive.org (consulté le 14 septembre 2012).

[12] Décret n° 2011-1904 du 19 décembre 2011 relatif au dépôt légal.

[13] Toute publication publiée ou diffusée en France est assujettie au dépôt légal.

- les sites enregistrés sous le nom de domaine .fr ou avec tout autre nom de domaine enregistré auprès des organismes français chargés de la gestion de ces noms[14] ;

- les sites dont le nom de domaine a été réservé par une personne physique ou morale domiciliée en France ;

- les sites proposant des contenus dont on peut prouver qu'ils ont été produits sur le territoire national (par exemple un blog hébergé aux États-Unis mais rédigé par un français).

Ainsi la mission de dépôt légal confiée à l'établissement permet de lever l'obstacle des difficultés d'ordre juridique. En effet, la mission de dépôt légal permet de se dispenser d'une demande d'autorisation auprès des ayants droits afin de copier et de diffuser des contenus publiés en ligne. Toutefois, la loi sur le dépôt légal de l'Internet diffère des précédents textes sur deux points essentiels. Premièrement, l'idéal d'exhaustivité est remplacé par un objectif de représentativité[15]. En effet, face à la masse de données que représente Internet[16], on tente de collecter un échantillon, le plus représentatif possible, de la production française sur le Web à un moment donné. Par exemple, les sites traitant des élections en 2012 sont représentatifs de ce que les Français ont produit et consulté à un moment donné. La seconde différence réside dans l'inversion des rôles entre le producteur et le dépositaire. En effet, l'obligation de dépôt légal n'implique pas de démarche particulière de la part des producteurs (sauf en cas de contenus protégés où le producteur doit fournir les codes et les informations techniques susceptibles de faciliter l'archivage de son site). Face à la masse de contenus présent sur la Toile et la multiplication du nombre de producteurs, il aurait été impossible de demander aux producteurs d'envoyer une copie de leur site à la Bibliothèque. Ainsi, c'est la BnF qui se charge de la collecte directement.

[14] L'AFNIC (Association française pour le nommage Internet en coopération) gère actuellement le registre des noms de domaine .fr (France), .re (Île de la Réunion), .yt (Mayotte), .wf (Wallis et Futuna), .tf (Terres Australes et Antarctiques), et .pm (Saint-Pierre et Miquelon).

[15] Voir Oury, Clément, « Soixante millions de fichiers pour un scrutin, les collections de sites politiques à la BnF », *Revue de la Bibliothèque nationale de France*, Paris, n° 40, 2012

[16] En septembre 2011, le nombre de noms de domaines enregistrés en .fr était d'environ 2,1 millions, ce chiffre ne représente que le tiers des enregistrements de l'AFNIC donc de « l'Internet français ». Source : AFNIC, *Observatoire du marché des noms de domaine en France - édition 2011*, p. 30-34. http://www.afnic.fr/medias/2011VF_observatoireWEB.pdf (consulté le 5 septembre 2012)

Dans la pratique, la collecte s'effectue par l'intermédiaire du robot Heritrix qui fonctionne comme un internaute automatique : partant d'une liste d'adresses URL[17], ce logiciel explore les sites Web concernés en naviguant de liens en liens soit en profondeur (liens entrant à l'intérieur d'un même site), soit en largeur (liens sortant vers d'autres sites), pour copier les fichiers qu'il rencontre: pages Web, images, sons, vidéos…[18] La liste d'URL de départ fixe également la profondeur et la fréquence de collecte.

La gestion et la coordination de cette collecte du Web sont assurées par le service du Dépôt légal numérique.

2. Le service du Dépôt légal numérique : interface entre bibliothécaires et ingénieurs

Au sein du département du Dépôt légal, le service du Dépôt légal numérique (DLN) compte cinq « chargés de techniques et de processus documentaires »[19] associés à quatre ingénieurs du département des Systèmes d'information. Le service collabore avec une équipe de sélectionneurs spécialisés[20], répartis dans les départements thématiques de la BnF. Ces derniers effectuent des sélections de sites selon leurs disciplines. Le service joue ainsi le rôle d'intermédiaire entre informaticiens et bibliothécaires en assurant la gestion et la coordination des collectes.

Au sein du département des Systèmes d'information (DSI), deux ingénieurs du service Support et production (SSP) et deux ingénieurs du service Études et développement (SED) sont entièrement mobilisés pour l'archivage du Web. Les ingénieurs SSP assurent le suivi des collectes au quotidien tandis que les ingénieurs SED sont chargés de développer les outils nécessaires à l'archivage du Web (logiciels permettant la sélection, la collecte ou encore l'accès aux collections).

Premièrement, le service du Dépôt légal numérique planifie et coordonne les collectes. En effet, les correspondants des départements thématiques de la DCO effectuent des sélections de

[17] L'URL (Uniform resource locator) représente l'emplacement de la page ou du fichier sur le Web.

[18] Voir Illien, Gildas, Oury, Clément, « Quelle politique documentaire pour l'archivage des sites Internet ? » dans Carbone, Pierre et Cavalier, François (dir.), *Les collections électroniques, une politique documentaire en mouvement*, Paris : Éditions du Cercle de la librairie, 2009, p. 157-178

[19] Anciennement nommés « chargés de collection numérique ».

[20] En septembre 2012, le nombre de sélectionneurs était de 80.

sites. À la réception de cette sélection, le DLN vérifie les paramètres de collecte et la structure des URL proposées[21]. Cette analyse des URL proposées à la collecte a un impact important sur la qualité de la collecte. Le service gère ensuite la planification et le lancement des collectes en collaboration avec les ingénieurs du DSI. Le calendrier de planification est essentiel pour s'assurer que les infrastructures et le personnel soient disponibles au lancement et pendant la durée de la collecte. Durant les phases de collecte, le service assure la surveillance en s'assurant de la qualité des contenus moissonnés[22]. Un contrôle qualité est également effectué en fin de collecte. L'objectif est d'améliorer les prochaines collectes et de corriger les problèmes rencontrés en cours de collecte. Ainsi, la BnF n'efface jamais une collecte effectuée même si cette dernière est incomplète ou non conforme à la demande des correspondants. Le contrôle qualité s'effectue sous la responsabilité des correspondants. Cette vérification se fait par l'intermédiaire de l'outil d'accès aux archives[23].

En résumé, voici la répartition des rôles entre les trois pôles[24] :

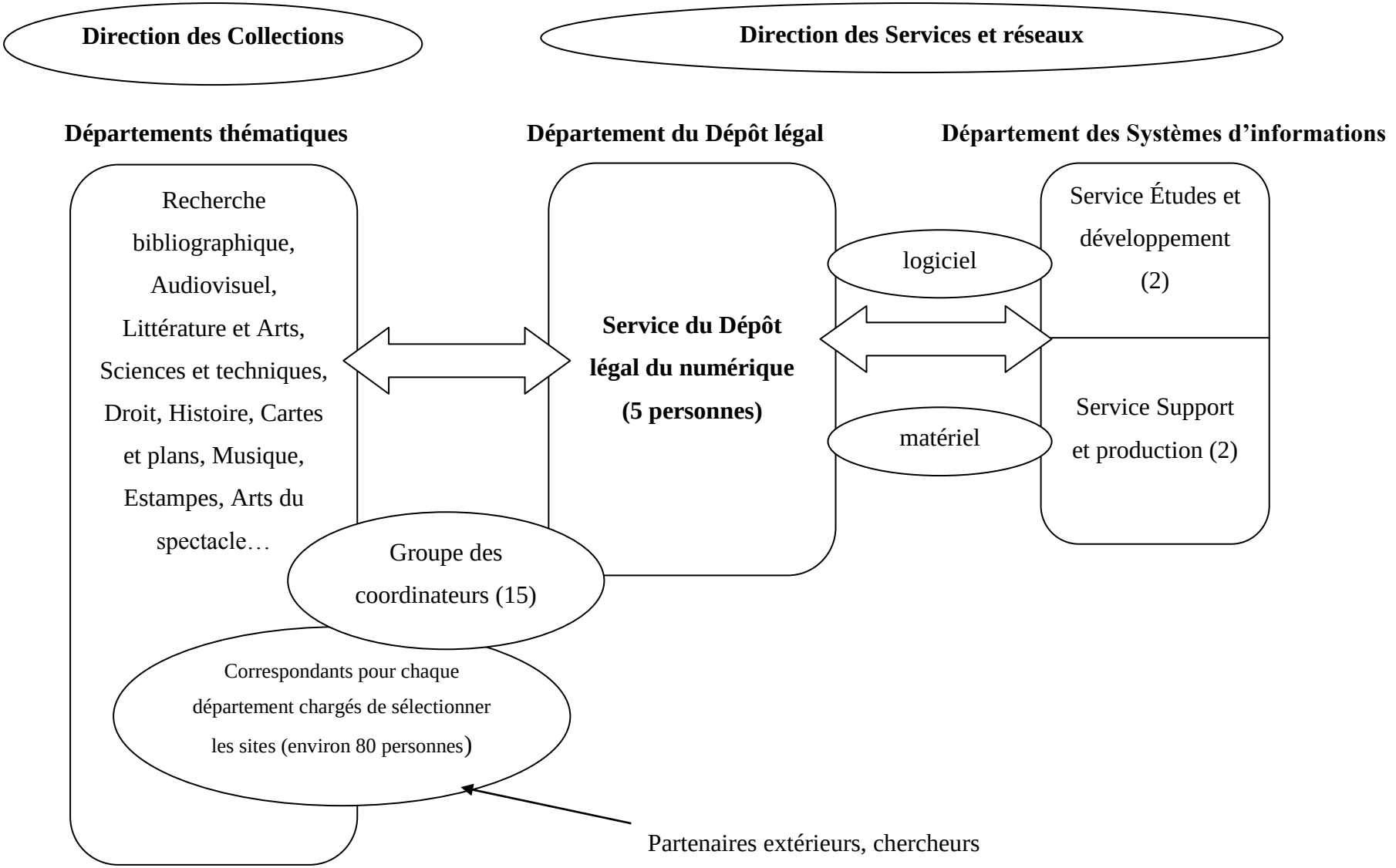
[21] Voir III. A. pour les détails concernant les paramètres de collecte.

[22] Voir III. B. pour les détails concernant les activités de surveillance.

[23] Il s'agit des Archives de l'Internet, outil basé sur la WayBack Machine, voir I. B. 4. et III. B. 2. pour plus d'informations.

[24] Voir III. pour obtenir plus d'informations sur les tâches effectuées par chaque entité.

Le service du Dépôt légal numérique, interface entre bibliothécaires et ingénieurs



L'équipe du DLN gère également les contacts avec les éditeurs ou producteurs des sites qui sont moissonnés par la BnF. Ceux-ci peuvent entrer en contact avec le DLN afin d'émettre une plainte, signaler un problème ou encore poser une question.

Le service du Dépôt légal numérique est aussi chargé de la formation des correspondants du dépôt légal du Web. En effet, une formation est nécessaire afin de maîtriser aussi bien les outils de sélection que la structure même du Web.

Enfin, le service du Dépôt légal numérique assure une activité de veille dans le domaine des technologies : évolution des formats, des architectures du Web… afin de se tenir à jour et d'adapter les outils de collecte en fonction des nouvelles évolutions. Le service est régulièrement amené à partager son expertise avec des établissements intéressés par l'archivage du Web. Le service collabore avec les bibliothèques de dépôt légal imprimeur en région, ainsi qu'avec les équipes de recherche, des universités ou encore des associations. Au niveau international, le service collabore avec de grandes institutions patrimoniales engagées dans l'archivage du Web dans le cadre du Consortium international pour la préservation de l'Internet[25].

Cette organisation de collecte des sites Web se structure autour de deux axes principaux, suivant les deux types de collecte : la collecte large et les collectes ciblées.

3. Deux types de collecte pour obtenir un contenu représentatif

Comme nous l'avons vu précédemment, il est impossible de satisfaire un idéal d'exhaustivité en archivant le Web. Les collections constituées se présentent donc plutôt comme des échantillons du Web liés entre eux comme dans leur environnement initial[26]. De plus, le Web est constamment alimenté par des flux d'information, puisque les sites sont mis à jour à une fréquence qui leur est propre. Ainsi, pour être véritablement exhaustif, il faudrait collecter en permanence les données. Les collections ainsi produites seraient énormes et ingérables. En fait, la technologie de la collecte du Web et l'espace de stockage ne permettent pas une telle opération. A partir là on peut envisager deux types d'approches pour obtenir une collection représentative de l'Internet français : la collecte large et la collecte ciblée.

[25] Voir I. C. 2.

[26] Illien, Gildas, « Le dépôt légal de l'Internet en pratique », *BBF*, 2008, n° 6, p. 20-27. Consultable en ligne : http://bbf.enssib.fr/ consulté le 4 septembre 2012

Les collectes larges sont mises en œuvre une fois par an, il s'agit d'une collecte peu profonde destinée à capturer l'intégralité du Web français[27] de manière superficielle à un instant donné. Ce type de collecte est très proche du principe du dépôt légal : des sites sont collectés sans aucun jugement sur leur intérêt documentaire. Il s'agit donc d'un échantillon suffisamment représentatif des contenus présents sur le Web français à un instant donné. La BnF mène cette collecte large en interne depuis 2010. La principale limite de cette approche réside dans le fait que des contenus considérés comme important ne peuvent être collectés que partiellement, alors que d'autres considérés comme ayant peu de valeur pour les chercheurs d'aujourd'hui représentent la majorité de ce qui est collecté (publicité, sites commerciaux, sites pornographiques…).

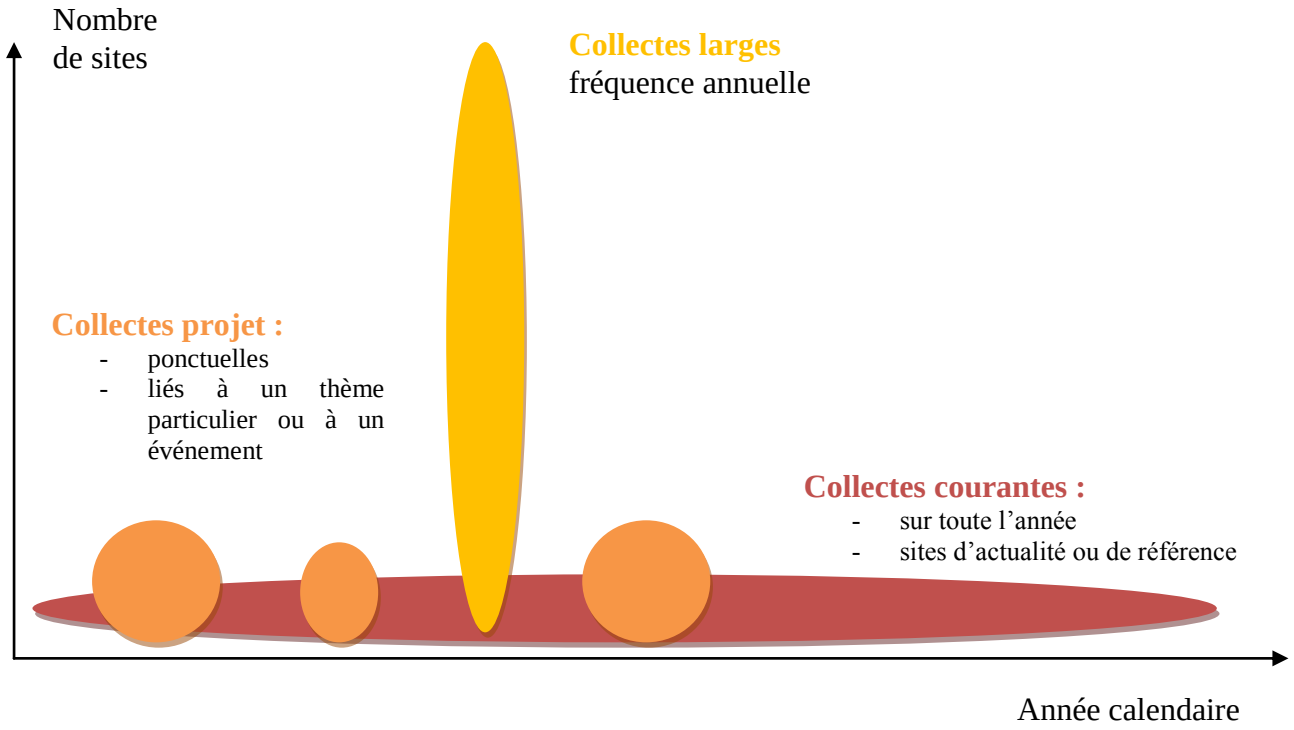
Quant aux collectes ciblées, ce sont des captures plus profondes réalisées à des fréquences plus rapprochées (quotidienne, hebdomadaire, mensuelle…) sur des sites choisis par les correspondants des départements thématiques suivant leurs disciplines ou bien selon des projets particuliers[28]. Cette option implique donc une sélection en amont, sur un critère de qualité ou de valeur scientifique. Ainsi des sites publiant de la recherche scientifique, des publications officielles, ou bien des travaux littéraires ou artistiques devraient être collectés en priorité. Ce type de collecte est très proche du système d'acquisition traditionnel des bibliothécaires. La principale limite de cette approche est le fait que la sélection requiert la définition de critères et un investissement de temps humain avec la possibilité que les sites sélectionnés aujourd'hui ne soient plus considérés comme les plus importants par les utilisateurs du futur.

La BnF a donc adopté un modèle intégré combinant à la fois collecte ciblée et collecte large pour obtenir un archivage du Web représentatif de l'Internet français à un instant donné.

[27] Il s'agit de la totalité des noms de domaine enregistrés en .fr auprès de l'AFNIC (2 millions en 2011) avec éventuellement d'autres sites sélectionnés par les correspondants

[28] C'est dans cette optique que sont réalisées les collectes électorales.

Le modèle intégré de la BnF :



Représentation tiré du document BnF *Le dépôt légal de l'Internet, Présentation aux nouveaux arrivants du DDL*

4. Conserver et communiquer les archives du Web

Une fois collectées, les archives (sous la forme de fichiers ARC[29]) sont d'une part stockées dans les baies NetApp qui sont des supports permettant l'accès aux données pour les lecteurs et d'autre part vouées à être sauvegardées dans SPAR (Système de Préservation et d'Archivage réparti)[30] afin d'assurer la conservation des archives sur le long terme.

La législation française permet pour le contenu collecté par le dépôt légal du Web, une mise à disposition similaire aux autres collections de dépôt légal, c'est-à-dire un accès strictement restreint. A titre de comparaison, d'autres pays comme le Royaume-Uni[31] ont un système

[29] Pour en savoir plus sur les formats ARC et WARC, consulter le site d'Internet Archive http://www.archive.org/Web/researcher/ArcFileFormat.php, consulté le 7 septembre 2012.

[30] Voir la page de présentation du projet SPAR sur le site de la BnF, à l'adresse http://www.bnf.fr/fr/professionnels/conserver_spar.html, consulté le 7 septembre 2012, ainsi que l'intervention de Bermès, Emanuelle, Dussert Carbon, Isabelle, Ledoux, Thomas et Ludovici, Catherine lors du congrès de l'IFLA en 2008, « La préservation numérique à la Bibliothèque nationale de France : présentation technique et organisationnelle », disponible à l'adresse : http://archive.ifla.org/IV/ifla74/papers/084-Bermes_Carbone_Ledoux_Lupovici-trans-fr.pdf, consulté le 7 septembre 2012.

[31] Plus d'information sur les archives du Web au Royaume-Uni et accès aux collections : http://www.Webarchive.org.uk/ukwa/, consulté le 7 septembre 2012.

d'archives en ligne : pour chaque site capturé, une demande de permission de collecte accompagnée d'une demande de rendre public le contenu, est adressée au producteur. La contrepartie est que cette obligation de demander la permission restreint considérablement le nombre de sites qu'il est possible de collecter. Internet Archive[32] fonctionne également avec des archives ouvertes, mais avec une politique basée sur « l'*opt-out* », les sites collectés sont mis en ligne et accessibles à tous, toutefois, en cas de plainte d'un producteur, Internet Archive retire les contenus concernés. Ces exemples montrent bien les difficultés auxquelles doivent faire face les institutions entre d'un côté la nécessité de l'accès et de la conservation, et de l'autre celui de respecter les droits de la propriété intellectuelle. Cependant, contrairement aux autres supports physiques, le fait d'utiliser une archive électronique ne l'endommage pas, puisque la loi autorise la reproduction des contenus à titre de conservation. C'est pour cela que la BnF sépare les copies utilisées pour l'accès et celles utilisées pour la préservation à long terme.

Ainsi, le Code du patrimoine spécifie que les contenus collectés au titre du dépôt légal numérique peuvent être consultés « sur place par des chercheurs dûment accrédités (…) sur des postes individuels de consultation dont l'usage est exclusivement réservé à ces chercheurs[33] » ; L'accès aux archives du Web est disponible dans les salles de recherche des différents sites de la BnF. Ces salles de lecture sont réservées à des chercheurs justifiant d'un besoin d'utiliser ces collections[34]. Cet accès retreint permet donc le respect du droit d'auteur et de la propriété intellectuelle mais aussi le respect de la vie privée, puisque la loi de 1978 sur l'informatique, les technologies et les libertés permet à des particuliers de modifier ou supprimer des informations les concernant contenus sur un site Web. L'existence d'un dépôt légal du Web suppose que des informations erronées sont accessibles dans les archives de la BnF d'où la nécessité d'un accès restreint et sous contrôle. Quant à la reproduction des contenus, elle est strictement limitée puisque seules les impressions d'écran sont autorisées.

L'interface d'accès aux collections des archives du Web de la BnF est basée sur la Wayback machine. Il n'existe pas, pour le moment, d'indexation plein texte des archives du Web. Toutefois, une indexation à titre expérimentale a été appliquée à environ 5 % de la collection.

[32] Accès aux collections d'Internet Archive : http://www.archive.org, consulté le 7 septembre 2012.

[33] Code du patrimoine, article L132-4 (*op. cit.*)

[34] Il ne s'agit pas uniquement d'un public universitaire, il est possible d'obtenir une accréditation si l'on souhaite consulter un document accessible uniquement à la BnF, ce qui peut être le cas avec les archives du Web.

La réalisation d'un moteur de recherche plein texte qui non seulement se calquerait sur nos habitudes d'utilisation du Web vivant, mais qui prendrait également en compte la dimension temporelle des archives, est un défi majeur pour l'avenir. Pour le moment, le principal moyen d'accès (en dehors des parcours guidés mis en place[35]) est une recherche par URL. Il faut donc connaître l'adresse exacte du site que l'on cherche dans les archives (l'URL exacte au moment précis où le site a été archivé).

Concernant la préservation des archives sur le long terme, les données sont stockées sur bande magnétique dans SPAR qui est un véritable magasin numérique. En effet, cet entrepôt de données sécurisées se charge de surveiller la présence et le bon état des contenus, d'en effectuer des copies, et devrait surtout garantir la continuité d'accès aux fichiers en réalisant les conversions nécessaires en cas d'obsolescence d'un format. Le fichier original est quant à lui toujours conservé.

C. Coopération nationale et internationale autour de l'archivage du Web

1. Les BDLI : partenaires au niveau national

Les bibliothèques du dépôt légal imprimeur (BDLI), une par région, sont habilitées à recevoir le dépôt légal des ouvrages imprimés dans leur circonscription. Il s'agit généralement des bibliothèques municipales des chefs-lieux de région[36]. Les BDLI participent au dépôt légal du Web en effectuant la même tâche que les correspondants des départements thématiques de la BnF, c'est-à-dire la sélection des sites qui méritent d'être collectés dans le cadre d'un projet donné[37]. En effet, certaines collectes ciblées comme les collectes électorales, nécessitent une participation humaine plus importante. À chaque collecte électorale, la BnF lance une proposition nationale de collaboration aux BDLI et s'associe à celles qui se portent volontaires[38]. Les BDLI opèrent un travail de sélection des sites, elles saisissent leurs

[35] Les « parcours guidés », accessibles depuis l'interface d'accès aux archives de l'Internet, présentent une sélection de sites sur un thème donné. L'objectif est de présenter la variété des ressources que la Bibliothèque a pu archiver. L'un des parcours guidés concerne justement les sites électoraux, de 2002 à 2007. Il montre ainsi une sélection de ressources illustrant la caricature politique sur le Web, la stratégie de communication Internet d'un candidat donné, ou encore les outils du militant sur la Toile.

[36] A noter que dans les régions d'Outre-mer, ce travail est assuré par les Archives départementales.

[37] Depuis 2004, à chaque collecte électorale, des BDLI ont participé à la sélection des sites.

[38] Le chiffre de BDLI volontaires a augmenté au fur et à mesure des collectes : elles étaient quatre en 2004 pour la collecte des élections régionales et européennes : les bibliothèques de Lyon, Caen, Rennes et Toulouse ; en

propositions de capture dans l'outil de sélection nommé BnF Collecte du Web (BCWeb)[39], outil accessible en ligne par identification. Le service du Dépôt légal numérique se charge de les traiter par la suite.

L'appel aux BDLI permet dans le cadre des collectes électorales, une meilleure couverture du territoire national et décharge les sélectionneurs de la BnF d'un travail de veille conséquent et difficile. En effet, les BDLI apportent un ancrage plus régional car elles disposent d'une meilleure connaissance de la vie politique locale et des acteurs locaux. Leur rôle est donc de sélectionner de la façon la plus complète possible une part « locale » du Web. La participation des BDLI est ainsi indispensable afin d'obtenir la collecte la plus représentatif de la campagne électorale, surtout lorsqu'il s'agit de scrutins locaux.

Par ailleurs, le décret d'application de la loi DADVSI prévoit d'ouvrir un accès contrôlé aux collections du dépôt légal de l'Internet dans un nombre limité de bibliothèques régionales (BDLI), pour la plupart déjà impliquées dans la sélection des sites.

D'autres partenaires à l'échelle nationale existent. Ponctuellement, dans le cadre d'une collecte projet, des équipes de recherche, des établissements universitaires ou encore des associations peuvent collaborer avec la BnF.

2. Un organisme international : International Internet Preservation Consortium (IIPC)[40]

A l'échelle internationale, la BnF travaille avec de grandes institutions patrimoniales engagées dans l'archivage du Web. Cette collaboration est coordonnée dans le cadre du International Internet Preservation Consortium (IIPC)[41]. Cet organisme a été fondé en 2003 à l'initiative d'une dizaine de bibliothèques nationales d'Amérique du Nord, d'Europe, dont la BnF, et de la fondation américaine Internet Archive. Les principaux objectifs d'IIPC sont la promotion de l'archivage du Web dans le monde, le développement collaboratif de logiciels libres et de normes pour la collecte, la formation du personnel, ainsi que la préservation et la

---

2007 pour la collecte des élections législatives, elles furent huit ; en 2010 pour les élections régionales, elles étaient au nombre de dix-neuf. Cette année pour les législatives, vingt BDLI ont participées.

[39] Pour le fonctionnement de l'outil, voir III. A. 1.

[40] IIPC : http://www.netpreserve.org/, consulté le 7 septembre 2012

[41] Illien, Gildas, « Une histoire politique de l'archivage du web », *BBF*, 2011, n° 2, p. 60-68
consultable en ligne : http://bbf.enssib.fr/, consulté le 07 septembre 2012

communication des archives de l'Internet. Depuis 2007, de nouvelles institutions ont rejoint le consortium[42]. Au niveau international, le consortium tente de sensibiliser les États aux questions liées à la préservation des contenus de l'Internet.

Cet organisme permet également aux institutions nationales de trouver des points de comparaison au niveau international puisque l'archivage du Web est une activité peu répandue à l'échelle d'un pays (échange de bonnes pratiques…). Chaque établissement a sa spécialité et peut la partager avec les autres membres. En matière de collecte, l'une des idées d'IIPC est que si chaque institution membre d'IIPC couvre son territoire national, idéalement, on pourrait aspirer à couvrir tout le Web mondial. Par exemple, dans le cadre de la collecte projet portant sur les Jeux Olympiques de 2012, un événement ayant une portée nationale, chaque institution a collecté les contenus Web de son pays.

La priorité du consortium réside dans le développement d'outils *open source* permettant l'archivage du Web. En outre, les logiciels utilisés par la BnF pour collecter le Web sont des logiciels libres développés de manière collaborative par les membres d'IIPC (chaque logiciel ayant été à l'origine développé par une seule institution) :

- Heritrix, logiciel de moissonnage[43] développé à l'origine par Internet Archive ;
- Wayback Machine, logiciel d'indexation et interface de consultation des archives[44] développé à l'origine par Internet Archive ;
- NutchWAX, logiciel expérimental d'indexation plein texte pour permettre la recherche par mot[45], développé à l'origine par Internet Archive ;
- Netarchivesuite, logiciel de gestion et de planification des collectes (https://netarkivet.statsbiblioteket.dk/suite/) développé à l'origine par NetArchive.dk (Bibliothèque royale du Danemark) ;

[42] Actuellement, IIPC regroupe une quarantaine de membres dont une grande majorité de Bibliothèques nationales européennes (outre la BnF, Espagne, Italie, Pologne, Royaume-Uni, Allemagne, Estonie, Danemark, Norvège, Suède, Finlande, République Tchèque, Slovaquie, Croatie, Suisse, Pays-Bas…), d'autres organismes européens (Institut national de l'Audiovisuel, Internet Memory Fondation…), de quatre institutions asiatiques (Japon, Singapore, Corée du Sud, Israël), d'une institution africaine (la Bibliothèque d'Alexandrie), des Bibliothèques nationales d'Australie et de Nouvelle-Zélande et des établissements nord américaine (Internet Archive, Bibliothèque du Congrès, Bibliothèque et Archives du Canada…).

[43] http://crawler.archive.org, consulté le 14 septembre 2012

[44] http://archive.access.sourceforge.net/projects/wayback, consulté le 14 septembre 2012

[45] http://archive-access.sourceforge.net/projects/nutch, consulté le 14 septembre 2012

- le format container WARC[46] a également été mis au point dans le cadre d'IIPC.

La BnF a récemment développé un logiciel qui permet de saisir les adresses des sites à collecter en leur associant un certain nombre de paramètres (fréquence de collecte, profondeur de collecte…) et une description documentaire. Cet outil nommé BnF Collecte du Web[47] organise les propositions des correspondants sous formes de fiches de site. Développé par la BnF, BCWeb pourrait voir son cercle d'utilisateur s'élargir à d'autres membres d'IIPC.

Concernant la gouvernance du consortium[48], les membres ne se rencontrent qu'une ou deux fois par an, son fonctionnement repose sur le volontariat. L'assemblée générale de l'organisme est composée d'un représentant pour chaque pays membre (une quarantaine de membres actuellement). Le comité de pilotage (*Steering committee*), composé de 15 membres (dont la BnF), vote les décisions importantes, notamment d'ordre budgétaire. Il se réunit deux ou trois fois par an. Quatre personnes, appelées officiers de l'IIPC, occupent des fonctions de coordination : le président (*Chair*), élu chaque année par le comité de pilotage, il représente l'ensemble de la communauté; le responsable de la communication (*Communication Officer*), désigné pour trois ans par le comité de pilotage; le responsable des programmes (*Program Officer*) est désigné dans les mêmes conditions, il coordonne les activités des groupes de travail ; enfin, le trésorier (*Treasurer*), désigné dans les mêmes conditions, assure la coordination financière du consortium (actuellement ce poste est occupé par le chef de service du dépôt légal numérique de la BnF). Trois groupes de travail complètent cette organisation. Chaque groupe de travail est piloté par un binôme issu de deux institutions différentes et est organisé autour des trois principaux thèmes : collecte, accès et préservation.

Il est intéressant d'observer cette organisation et de comprendre que l'archivage du Web ne peut être effectué efficacement qu'à l'échelle de la planète.

[46] Le format container WARC est l'héritier du format ARC actuellement utilisé par la BnF. Les données collectées sur le Web sont regroupées au sein de fichier container ARC. Un fichier ARC regroupe plusieurs enregistrements ARC. Un enregistrement ARC est le fichier collecté sur le Web, accompagné de métadonnées.

[47] Voir le fonctionnement de l'outil en III.

[48] Voir organigramme en annexe.

II. Etude de cas : la collecte des sites Web lors des campagnes électorales de 2012

L'archivage des sites pour les élections présidentielle et législative de 2012 vient compléter une collection de sources primaires constituée par la BnF sur une période de dix ans. En effet, la collecte du Web électoral de 2012 représente une suite logique des opérations d'archivage effectuées par la BnF à l'occasion des élections de 2002 (présidentielle et législatives), 2004 (régionales et européennes), 2007 (présidentielle et législatives), 2009 (européennes) et 2010 (régionales).

À travers cette collecte projet, nous allons voir l'intérêt de constituer une telle collection, détailler les objectifs, décrire la collection constituée et expliquer le processus de sélection des sites à archiver.

A. Pourquoi vouloir archiver ces sites en particulier ?

1. Le rôle croissant des Web campagnes

Après les campagnes électorales de 2007, celles de 2012 étaient annoncées comme les secondes « Net-campagnes » de la France. Certes lors des scrutins précédents, les discussions politiques et les débats électoraux s'étaient transposés sur le Web, toutefois ce n'est que depuis 2007 que les candidats ont investi à part entière l'espace Internet. Lors du référendum de 2005, le Web avait été le média privilégié des partisans du « non » au Traité Constitutionnel Européen. Plus récemment, la campagne de Barack Obama pour l'élection présidentielle aux États-Unis, a démontré que le Web pouvait jouer un rôle majeur dans l'organisation de la campagne.

L'influence croissante du Web politique s'appuie d'abord sur des audiences désormais bien établies. En effet, en mars 2012, près de deux Français sur trois pouvaient être considéré comme des internautes selon Médiamétrie[49]. Internet fait donc partie intégrante de la vie politique française.

Durant la campagne électorale, l'Internet a souvent été perçu comme un média complémentaire de la télévision. En effet, cette année encore, une part importante de la

[49] 40 millions de français exactement. Un internaute est défini comme un Français de plus de 11 ans s'étant connecté à l'Internet au moins une fois au cours du dernier mois, quel que soit son mode de connexion. Source : http://www.mediametrie.fr/internet/communiques/hotspot-internet-le-web-social-live-et-video.php?id=625, consulté le 8 septembre 2012.

campagne s'est déroulée sur les plateaux de télévision. Mais contrairement aux autres années, Internet a cette fois-ci joué un rôle dans la campagne télévisée par l'intermédiaire du *live-tweeting* et du *fact-checking*. En effet, la pratique du *live-tweeting*[50] a particulièrement été appréciée lors des émissions telle que le débat de l'entre deux-tours, ainsi le hashtag[51] #Ledebat était parmi les plus utilisés du monde avec près de 1 500 tweets à la minute le soir du 2 mai[52]. Quant au *fact-checking*, il s'agit là encore d'une intervention en direct sur le Web afin de vérifier les paroles d'un homme politique sur un plateau de télé (contradiction, confirmation ou infirmation des chiffres donnés…). La plupart des sites d'informations (Le Monde, Le Figaro, le Huffington Post…) ont proposé ce type d'éclaircissement en temps réel à leurs internautes. D'autres sites ont également effectué un travail similaire en adoptant un recul par rapport aux événements (le véritomètre Owni[53] ou encore le désintox de Libération[54]).

Par ailleurs, l'Internet a aussi été un média proposant du contenu et des informations qui lui étaient propres. Outre les sites et réseaux sociaux des candidats, de nombreux sites comparateurs de programmes se sont multipliés pendant la campagne (http://jevotequien2012.fr,…).

Les partis politiques ont désormais eux aussi bien investi la Toile. Les candidats ont pris conscience de l'importance de ce média : l'Internet offre une possibilité de présence pour tous, même les plus petits, sans le déficit du temps de parole. Des moyens humains et financiers sans précédent ont été consacrés à la Web campagne par les partis politiques[55]. L'Internet permettait donc aux candidats et aux médias traditionnels d'envoyer un flux continu d'information vers les électeurs (déplacements des candidats, meetings…). Déjà en

[50] Live-twitter signifie envoyer des messages sur Twitter à propos d'un événement auquel on est en train de participer. Twitter étant un outil de réseau social qui permet à un utilisateur d'envoyer des messages (appelé Tweet) limités à 140 caractères.

[51] Sur les réseaux sociaux, le hashtag sert à centraliser les messages autour d'un terme bien précis. Il fait office de mot-clé pour que les utilisateurs puissent commenter ou suivre une conversation.J'ai découvert qu'en français, on disait « mot-clic »

[52] Source : http://presidentielle2012.ouest-france.fr/actualite/sur-twitter-refait-le-debat-03-05-2012-1477...

[53] http://itele.owni.fr/, consulté le 11 septembre 2012.

[54] http://desintox.blogs.liberation.fr/, consulté le 11 septembre 2012.

[55] En moyenne, environ 10 % du budget total alloué à la campagne de chaque parti a été consacré à la Web campagne. Source : *Journal Officiel* du 31 juillet 2012, accessible en ligne : http://www.legifrance.gouv.fr, consulté le 8 septembre 2012.

2007, les militants se sont emparés du Web en construisant eux-mêmes leur propre réseau avec pour objectif de commenter l'actualité de la vie politique, de se moquer ou au contraire de mettre en valeur une personnalité politique, de soutenir un parti, de mobiliser les citoyens autour d'une action politique… En 2012, les partis politiques et leurs candidats ont également investi le terrain des réseaux sociaux. En effet, tous les candidats à l'élection présidentielle et une grande partie des candidats aux législatives disposaient d'un compte Facebook et d'un compte Twitter.

Enfin, le Web a également était un espace de parodie et de vision décalée de la campagne présidentielle. Cette vision décalée passe par les comptes Twitter des politiques (exemple les « DM Fail »[56] de Nadine Morano, polémiques autour du tweet de Valérie Trierweiller) mais également par les animations créées pour l'occasion (http://fh-2012.com[57]) ou encore par les très nombreux montages photos (lesrésultats.tumblr.fr, ledebat.tumblr.fr, …). Quant au hashtag #radioLondres, il a permis aux internautes de communiquer les résultats des scrutins avant 20 heures sans risquer l'amende[58]. Il semble que même les partis politiques voient le Web comme un outil permettant une vision décalée des élections, puisque l'équipe de campagne de François Bayrou a proposé sur son site officiel une animation type jeu vidéo dissimulée par un code[59].

Ainsi, les contenus Web de la campagne électorale de 2012 ont joué un rôle qui leur est propre au même titre que les autres médias. Si l'on ne sauvegardait pas ces contenus, on perdrait toute une partie de la campagne même. De plus, bien que très actifs le temps de la campagne, ces contenus Web risquent de disparaître assez rapidement une fois les élections terminées.

[56] DM fail : sur Twitter, signifie échec d'un message privé, il s'agit d'un message privé destiné à une personne donnée qui est posté par erreur dans le domaine public.

[57] Ce site propose d'enfariner François Hollande, faisant référence à un événement de la campagne.

[58] Etant donné l'interdiction de divulguer les résultats avant 20 heures, les internautes ont communiqués les résultats en utilisant des phrases codées. A noter que les estimations étaient également disponibles sur les sites belges et suisses à partir de 18 heures, comme les années précédentes.

[59] Il s'agit d'un Konami Code (utilisé dans les jeux vidéo édités par Konami). Appliqué depuis à de nombreux sites Internet, le code permet de déverrouiller les contenus cachés d'un site. Le code consiste à taper la séquence suivante : haut, haut, bas, bas, gauche, droite, gauche, droite, b, a.

2. Des contenus éphémères ?

Si depuis 2007, le rôle joué par les Web campagnes progresse, ces contenus sont souvent voués à une courte durée de vie, se limitant dans la majorité des cas au temps de la campagne. En effet, si Internet permet de produire et de diffuser des contenus avec beaucoup de facilités, la durée de vie de ces contenus en ligne est relativement faible. Dans le cadre d'une campagne électorale, les contenus mis en ligne ont pour objectif de peser sur le scrutin. Une fois les élections terminées, conserver en ligne les données pourrait même s'avérer contre-productif. En effet, le discours présent en ligne pourrait être un frein à de nouvelles alliances politiques par exemple. En 2007, une étude menée par la bibliothèque municipale de Lyon a montré que plus de la moitié des sites qu'ils avaient identifié pour les législatives en Rhône-Alpes avaient disparu cinq mois plus tard[60]. La grande majorité des sites qui disparaissent rapidement sont des sites de candidats aux législatives, qui ferment au soir du premier tour des élections, afin de ne pas mettre en péril une nouvelle liste d'union.

De manière générale, les contenus en ligne ont un caractère relativement éphémère. En effet, comme nous l'avons vu précédemment, il suffit que l'auteur d'un site se désintéresse ou même décide de sa fermeture pour que les contenus disparaissent. Par ailleurs, si l'auteur d'un site oublie de renouveler la redevance annuelle pour son nom de domaine, c'est-à-dire l'adresse où se situe le site, il y a également un risque de disparition. De même, il se peut que l'hébergeur technique du site rencontre des problèmes pour que des liens soient cassés et empêchent d'accéder aux contenus du site. D'autre part, de nombreux sites ne conservent pas de traces de leurs archives traitant de la période électorale.

L'expérience des collectes électorales menées par la BnF précédemment permet justement de se rendre compte de l'ampleur du risque de disparition de ces sites. Lors des campagnes de collecte électorale de 2002 et 2004, un total de 3 000 sites ou parties de sites avaient été capturés. Dès juin 2006, le tiers de ce qui avait été collecté n'existait plus en ligne. En août 2007, plus de 6 % des sites capturés concernant l'élection présidentielle avaient disparu en l'espace de quelques mois[61]. Il a également été remarqué que le taux de disparition était plus

[60] « La netcampagne des législatives 2007 en Rhône-Alpes : la course au Net et après », *Points d'actu !* , Lyon, bibliothèque municipale de Lyon, 23 juillet 2009, disponible en ligne : http://www.pointsdactu.org/article.php3?id_article=863, consulté le 9 septembre 2012.

[61] Il s'agit d'un calcul effectué de façon automatique, par un outil interrogeant les sites en ligne, et qui note les codes de réponses. Le pourcentage de perte correspond au taux de réponses 404 (*File Not Found*). Le taux de disparition peut en fait être plus fort que ce qu'indique l'outil. En effet, l'outil considère comme positive une

élevé pour les sites de personnalités pressenties pour être candidates, mais qui ne le sont pas finalement. Si l'on regarde les sites des candidats à l'élection présidentielle de 2012, trois des candidats n'ont plus leurs sites de campagne accessibles en ligne[62].

Ainsi, Internet est devenu un véritable outil de campagne, l'intensité des Web campagnes va croissant depuis 2007. Toutefois, les contenus sont souvent voués à une courte durée de vie. Les établissements en charge de leur collecte doivent donc mettre en place un modèle de collecte adapté.

B. Comment les archiver ?

1. Mise en place d'un modèle documentaire

Les collectes projet autour des élections se répètent régulièrement, et la BnF a mis au point depuis 2007 une typologie documentaire permettant d'encadrer la sélection des sites. Cette typologie devait permettre de prendre en considération tous les acteurs politiques et leurs caractéristiques et d'ordonner au mieux cette collecte thématique. La typologie devait également permettre de saisir toutes les nuances de l'opinion publique politique.

Pour atteindre ces objectifs, la BnF a fait le choix d'une typologie qui s'intéressait non pas à la forme des sites, c'est-à-dire à leur modèle de publication, mais à leur fond. Par exemple, le blog ne constituait pas une rubrique en soi. Le critère principal a été l'identification de l'émetteur du site. Il s'agissait d'identifier sa logique de positionnement, en tant qu'acteur, dans l'espace public virtuel.

Ainsi, la sélection s'est concentrée sur deux axes. D'abord, on distingue les candidats et leurs organisations, c'est-à-dire « l'offre politique » proposée aux électeurs internautes dans le cadre de ces scrutins. Pour cette première partie, qui s'inscrit dans la continuité des précédentes collectes, le cadre de la sélection devait être assez systématique et donc être le plus complet possible. Autour de tous les candidats, et des différentes formations qui contribuent à leur stratégie électorale, les sites des partis politiques, mais aussi les sites et blogs officiels ou personnels des candidats en campagne ont été sélectionnés. La seconde partie de la typologie aborde les regards et opinions sur la campagne ; cet axe apporte un

réponse 200 (« requête traitée avec succès »), alors que la page renvoyée peut être vide. Source : document BnF « Elections 2007 : Bilan documentaire du projet ».

[62] Il s'agit des sites de campagne de Philippe Poutou, de Nicolas Sarkozy et de François Hollande, les deux premiers renvoient vers les sites des partis respectifs tandis que le troisième renvoie vers la page Facebook de François Hollande.

ensemble d'illustrations, de réactions et de contributions aux débats citoyens. Cette partie devait permettre de retrouver sur le Web des acteurs collectifs des campagnes (comités de soutien, syndicats, autres organisations constituées intervenant à des titres divers dans le débat public), mais aussi d'explorer des formes émergentes d'expression individuelle ou communautaire. Compte tenu de la profusion et de la volatilité des sites dans ce domaine, les sélectionneurs de la BnF et des BDLI ont tenté de proposer un échantillon de sites aussi représentatifs que possible.

Ainsi, la typologie en vigueur depuis 2007 pour les collectes électorales est la suivante :

0 – **Sites officiels et institutionnels** : règles du jeu et cadre juridique (ministère de l'intérieur, CSA, collectivités).

1 - **Les candidats et leurs organisations**

1.1 Sites des candidats en campagne : sites et blogs officiels ou personnels, permanents ou de campagne, des candidats

1.2 Sites des formations politiques : partis et formations en campagne

1.3 Autres organisations de soutien : uniquement les sites clairement affiliés aux partis et aux candidats : comités de soutien par exemple

2 - **Regards et opinions sur la campagne**

2.1 Annuaires, observatoires et analyses : y compris les sites de sondages et de tendances, les nouveaux outils d'analyse de type cartographique, les sites de fact-checking…

2.2 Médias traditionnels : sélection restreinte à un nombre limité de titres de presse d'information disposant d'un site à l'activité significative, seul la partie consacrée aux élections des grands quotidiens a été sélectionné.

2.3 Associations, syndicats et autres organisations : cette catégorie servait à relayer la parole d'acteurs collectifs qui n'existent pas uniquement sur l'Internet.

2.4 Expressions individuelles et communautaires sur l'Internet : cette catégorie vise à regrouper les formes d'expression et d'agrégation de l'opinion de la société civile (blogs…)

D'autre part, chaque site sélectionné était associé à un certain nombre de mots-clés dans l'outil de sélection BnF Collecte du Web. Il a été décidé d'utiliser un vocabulaire normé et une présentation uniformisée pour ces mots-clés :

- Un premier niveau obligatoire correspond au niveau géographique : les sélectionneurs devaient indiquer la région concernée par la proposition (ou national dans le cas de l'élection présidentielle).

- Un deuxième niveau correspond à l'appartenance partisane (si pertinent) : ce niveau est renseigné dans la plupart des cas (noms des partis écrits en abrégé selon une nomenclature prédéfinie).

- Un troisième niveau correspond à la personne (si pertinent): ce niveau désigne les personnalités concernées par le site : souvent l'auteur ou le sujet principal du site.

- Les niveaux suivants sont facultatifs et libres : ils peuvent renseigner un département, une ville, un thème abordé par le site…

Ces mots-clés permettent d'effectuer des tris et de connaître la répartition du nombre de sites en fonction de la zone géographique ou du parti politique.

Une fois la typologie et les mots-clés définis, il fallait mettre au point des critères de sélection communs à tous les sélectionneurs afin d'obtenir la sélection la plus homogène et la plus représentative possible.

2. Repérage et sélection des sites

La sélection des sites reposait donc sur un certain nombre de critères prédéfinis. Bien que la sélection dût être large et ouverte, il fallait respecter les périmètres du dépôt légal numérique et ceux plus spécifiques de la collecte électorale. Les critères suivants devaient être respectés[63] :

- critère **légal** du périmètre de collecte de la BnF : les sites de radio et de télévision sont du ressort de l'INA ;

- critère **géographique** : les sites sélectionnés sont ceux dont les auteurs, contributeurs ou éditeurs sont localisés en France ou de nationalité française, s'exprimant en

[63] Ces critères ont été repris du document BnF intitulé «Le dépôt légal de l'Internet électoral en 2012. Journée de travail et de formation avec les BDLI partenaires ».

français ou dans une langue régionale ; les regards portés sur la campagne depuis l'étranger (même s'ils sont francophones ou de grande réputation) sont donc exclus du périmètre ;

- critère de **contenu actualisé** : les sites ou parties de sites retenus doivent traiter de la campagne de 2012, et être régulièrement mis à jour ;

- critère de **singularité** et d'**originalité** : les sites doivent proposer un contenu signifiant et original (il ne s'agit pas de collecter des sites se contentant de reprendre, sans les remettre en forme ou en perspective, des contenus déjà accessibles par ailleurs) ;

- critère de **représentativité** : il faut veiller autant que possible à un équilibre idéologique entre les différents courants de pensée qui s'expriment : il ne s'agit pas de chercher à tout prix des équivalences qui n'existent pas en ligne, mais de respecter le pluralisme politique en couvrant la diversité des débats, et en représentant toutes les grandes tendances qui s'expriment sur l'Internet ; par ailleurs, l'offre sélectionnée doit rester conditionnée à la **publicité** de la source : le contenu soumis à identification des comptes sur les réseaux sociaux, la correspondance privée ou les intranets ne sont pas concernés ;

- critère de **navigabilité :** il s'agit de repérer le site ou partie de site le plus significatif, et de ne pas collecter des documents non pertinents (cette règle est surtout valable pour des sites généralistes, comme la presse ou les sites de sondage). Néanmoins, même si l'outil de saisie permet d'aller au plus précis dans le choix de la capture, ne pas oublier que l'unité documentaire de référence demeure le site. Il faut toujours penser aux façons dont les futurs lecteurs pourront naviguer dans les archives. Ainsi, plutôt que de multiplier la collecte de documents à la page dans un site, il est souvent plus satisfaisant de proposer une capture au niveau supérieur, qui permettra de contextualiser la publication et de la relier à d'autres.

Concernant le repérage des sites, chaque sélectionneur a utilisé ses propres sources, en fonction de son domaine de compétence. Toutefois, plusieurs outils communs pour effectuer la veille, ont été utilisés[64].

[64] Voir III. A. 2. pour avoir plus de détails sur ces outils.

Une fois que cette méthode documentaire d'archivage a été assimilée, on peut se questionner sur les objectifs d'une telle collecte, en d'autres termes on peut se demander à qui s'adresse les collectes du Web électoral.

C. Quels sont les objectifs ? Quel public vise-t-on ?

1. Constituer une mémoire de l'Internet politique

Comme nous l'avons vu précédemment, la documentation politique s'est désormais bien implantée sur le Web. Les nombreuses possibilités offertes par les partis politiques aux militants, telle que l'impression de tracts, nous amènent même à nous interroger sur une éventuelle mutation de la documentation politique vers le numérique. L'utilisation du Web en politique, c'est aussi l'idée d'une démocratie plus directe, l'idée que l'on s'adresse au citoyen, chez lui sans autres intermédiaires. Par exemple, en 2007, Ségolène Royal, candidate à l'élection présidentielle, avait lancé un site participatif, encourageant à l'action militante[65]. Cet exemple montre à quel point l'Internet politique est mouvant et éphémère. En effet, le Web possède une temporalité qui lui est propre, c'est en quelque sorte un flux d'information en continu. Cette temporalité entre en contradiction avec le discours politique, il est de l'intérêt d'un parti, qu'un discours officiel et uniforme demeure. La pluralité des acteurs sur le Web empêche cette uniformité. C'est notamment pourquoi les contenus du Web politique sont de nature éphémères.

L'archivage des réseaux sociaux (Facebook, Twitter…) pose problème dans la mesure où ces contenus ne sont pas destinés à devenir pérennes puisqu'ils relèvent du domaine de la communication. Quel sens doit-on donner à un message posté sur un de ces réseaux sur le long terme ? Il est donc nécessaire d'effectuer un travail de recontextualisation à partir des données archivées.

La dimension parodique du politique s'est également largement implantée sur le Web. On peut même parler d'une surreprésentation des détournements politiques sur l'Internet. Ces

[65] http://www.desirsdavenir.org a été retiré du Web à la suite de l'échec électoral de 2007. En 2009, le site a été relancé sous une autre forme, consultée le 10 septembre 2012.

contenus souvent présentés sous forme de blogs sont des zones encore plus sensibles à la volatilité (arrêt volontaire de l'auteur, condamnation et fermeture du site[66]…).

L'archivage des sites institutionnels (sites des ministères, sites de l'Elysée, des administrations et établissements publics…) apparaît comme plus facile à délimiter et à collecter car ce sont des contenus qui nous apparaissent plus stables. D'un point de vue citoyen, il nous parait évident d'archiver ces sites en priorité. En fait, l'archivage de ces contenus permet de remettre en question la stabilité apparente de ces sites. Il est possible de montrer comment les discours officiels ont évolué, se sont contredits ou ont été corrigés sans prévenir. D'ailleurs, lors de la prise de fonction de François Hollande en tant que Président de la République, le site de l'Elysée a été entièrement vidé de ses contenus[67]. L'archivage des sites publics permet donc d'exercer, en quelque sorte une surveillance citoyenne.

Par la nature de leur environnement, les contenus du Web sont toujours mouvants. Ajouts, suppressions, modifications, les sites sont constamment retouchés aussi bien sur le fond (les contenus) que sur la forme (aspect visuel du site). Cette particularité est d'autant plus vraie pour les contenus politiques qui sont désormais bien implantés sur le Web. Il est donc nécessaire de constituer une mémoire de l'Internet politique. En effet, l'archive d'un site Web est un moyen de mettre de la distance à l'égard de l'objet, de le figer pour mieux l'étudier. Cependant, si les chercheurs s'accordent sur le fait qu'il est nécessaire de conserver une mémoire du Web, peu utilisent ces archives du Web dans leurs travaux pour le moment. En effet, il existe une certaine réticence envers ces contenus, probablement dûe au fait qu'il n'existe pas de corpus aux contours maîtrisés. Toutefois, ces archives s'adressent à un public assez large qui va de l'usager citoyen jusqu'au chercheur en Web politique.

2. De l'usage citoyen au chercheur en Web politique

L'accès aux collections du dépôt légal numérique se fait uniquement dans les espaces recherche de la BnF. Tout usager qui peut justifier de sa recherche, simple citoyen comme chercheur, peut y accéder.

[66] Durant la campagne pour les législatives de 2012, Patrick Balkany, candidat à sa réélection dans les Hauts-de-Seine, a porté plainte contre un site le parodiant. Le site en question est toujours en ligne à la date du 10 septembre 2012 : http://voteinutile.fr/patrick-balkany.

[67] http://www.elysee.fr a fait totalement peau neuve. Les contenus relatifs à la présidence de Nicolas Sarkozy sont désormais en ligne, de manière temporaire, à cette adresse : http://www.archives.elysee.fr/. De la même manière, les sites de Matignon et des Ministères ont également été vidé de leurs contenus, mais cette fois-ci, il n'y a pas d'archives disponibles en ligne.

Dans l'idée d'exercer une surveillance citoyenne, l'élu doit rendre des comptes, non seulement à ses pairs, mais également à tout citoyen. Le Web est un espace privilégié pour l'exercice de ce contrôle citoyen. Il permet au citoyen d'être de mieux en mieux informé. Ce dernier peut par exemple, confronter les réalisations d'un élu par rapport aux engagements qu'il avait tenu avant son élection. Les archives permettent de remonter à ces engagements par l'intermédiaire du site officiel du candidat, puis de suivre ces engagements au fil du temps et des déclarations qui suivent l'élection. Les chercheurs amateurs et les étudiants sont également un public visé. D'autres collections du dépôt légal numérique (sites d'entreprises…) peuvent susciter l'intérêt de professionnels tels que des avocats, des journalistes, des ingénieurs brevet ou encore des responsables d'entreprises. Cet usage potentiel est majoritairement celui de la preuve.

D'autre part, le premier usage qui nous vient à l'esprit lorsqu'on parle d'archives des sites Web traitant du politique, ce sont les travaux liés à l'analyse du Web politique. En France, ce terrain est investit notamment par Fabienne Greffet[68], maître de conférences en sciences politiques à l'Université Nancy II ; Dominique Cardon[69], sociologue au laboratoire des usages d'Orange Lab ; Frank Rebillard, maître de conférences en sciences de l'information et de la communication, à l'Université Lyon II ou encore Thierry Vedel[70], chargé de recherche CNRS au CEVIPOF.

Cependant, un certain nombre de représentations du Web viennent freiner encore aujourd'hui, l'intérêt du public pour les archives du Web[71]. Premièrement, il s'agit de l'idée d'un usage de l'Internet essentiellement tourné vers l'instant présent, il est donc difficile d'associer le Web avec l'idée d'archive. D'autre part, la masse d'information présente sur Internet est tellement importante qu'elle semble se suffire à elle-même. C'est l'idée qu'il y a déjà suffisamment de choses à traiter sur le Web vivant alors pourquoi recourir aux archives ? De plus, cette masse d'information est tellement énorme qu'elle semble constituer sa propre archive. Cette idée est

[68] Parmi ses travaux, on peut citer : Greffet, Fabienne (dir.), *Continuerlalutte.com. Les partis politiques sur le web*, Paris, Les presses Sciences Po, 2011 ; "Le web dans la recherche en science politique. nouveaux terrains, nouveaux enjeux", *Revue de la BnF*, Paris, n° 40, p. 78-83.

[69] Cardon, Dominique, *La démocratie internet. Promesses et limites*, Paris, Seuil, 2010.

[70] Il a notamment contribué à l'ouvrage collectif publié sous la direction de Fabienne Graffet, cité précédemment (*Continuerlalutte.com.)*

[71] Voir à ce sujet Chevallier, Philippe, Illien, Gildas, *Les Archives de l'internet. Une étude prospective sur les représentations et les attentes des utilisateurs potentiels*, Paris, BnF, 2011, en ligne : http://www.bnf.fr/documents/enquete_archives_web.pdf, (consulté le 10 septembre 2012)

renforcée par la manière dont on se sert des moteurs de recherche tels que Google, où il est possible de trier sa recherche par date de mise en ligne des contenus.

Après avoir vu les objectifs de la collecte du Web politique et le spectre des différents publics possibles pour ces archives, nous allons décrire la collection constituée du Web électoral de 2012.

D. La collection constituée du Web électoral de 2012

1. Volumétrie de la collection

Le travail de repérage, de sélection, de collecte et d'archivage, a permis de constituer une collection du Web électoral de 2012. Cependant la diversité des documents recueillis rendent cette collection difficile à appréhender. Pour avoir une vision d'ensemble de la collection afin de la comparer avec d'autres collections du même ordre, il est nécessaire d'en établir la volumétrie.

Afin d'y parvenir, il est indispensable dans un premier temps de définir des unités de mesure. Sur la Toile, l'unité la plus évidente semble être celle du site Web. Pourtant, cette unité éditoriale est difficile à appréhender techniquement. Si on conçoit bien ce qu'est le site Web de l'Assemblée nationale ou du journal *Le Monde*, peut-on considérer qu'un blog de la plateforme Over-blog est un site à part entière ? Le site, n'est-ce pas plutôt la plateforme elle-même ? D'autre part, le site d'Europe Écologie les Verts (http://www.eelv.fr) héberge un très grand nombre de sites des sections locales du parti (du type http://ville.eelv.fr). Dans ce cas qu'est-ce qui est considéré comme site ? L'ensemble ou bien chacune des parties ?

En fait, le Web a un modèle éditorial où les contenus s'emboîtent les uns dans les autres. Ainsi au niveau le plus haut, le Web est découpé en *Top Level Domains* (domaines de haut niveau). Ces derniers sont génériques (.com, .org, .net, …) ou nationaux (.fr, .us, .uk, …). On peut y acheter des noms de domaines, par exemple, EELV a acheté le nom de domaine eelv.fr. Ces noms de domaine peuvent être divisés par leur propriétaire en autant d'hôtes que besoin (angouleme.eelv.fr, bourgogne.eelv.fr,…). Ces hôtes correspondent souvent aux supports physiques (serveurs) qui hébergent les fichiers. Les hôtes peuvent à leur tour être divisés en dossiers (angouleme.eelv.fr/category/actualite/). Enfin au niveau le plus bas, on trouve la page Web qui est un fichier contenant d'autres fichiers encapsulés (images, sons).

Ainsi, il est difficile de définir une unité de référence. Toutefois, pour estimer le volume de la collection du Web électoral de 2012, on peut s'intéresser à son point de départ : les URL sélectionnées. En effet, le travail de sélection effectué par les agents de la BnF et des BDLI a permis d'identifier de capturer un grand nombre de sites Web (ou des parties de sites). Au total, plus de 10 500 sites[72] ont été sélectionnés pour la collecte du Web électoral de 2012 ; à titre de comparaison 1 900 sites avaient été sélectionnés en 2002 et 5 800 l'avaient été pour 2007. Parmi ces sites, 40 % ont été sélectionnés par les correspondants de la BnF tandis que le reste l'a été par les BDLI participants au projet. Cette part de la BnF s'explique par le fait que les correspondants de la BnF ont sélectionné les sites de niveau national, les sites pour l'Île-de-France et les sites des régions pour lesquelles les BDLI n'ont pas été partie prenante. Toutefois, comme nous l'avons indiqué précédemment, ce chiffre de 10 500 sites sélectionnés ne suffit pas pour autant à caractériser la collection ainsi constituée. Puisqu'il ne s'agit pas de 10 500 « sites Internet » à proprement parler, mais des adresses URL à partir desquelles le robot, en fonction des paramètres de profondeur qu'on leur attribue, peut parcourir le Web.

Sur six mois de collecte (janvier à juin), ce sont près de 300 millions d'URL[73] qui ont été collectées, ce qui représente un poids d'environ 9 156 Go de données. Sachant que les données collectées au titre du dépôt légal du Web représentent environ 80 To par an, la collecte du Web électoral de 2012 représente à elle seule, plus de 11 % de la production annuelle.

Afin de mieux caractériser cette collection numérique, il est possible de connaître le format des fichiers archivés. Plus de 70 % des fichiers archivés sont au format html, il s'agit donc de texte. Les différents formats image (jpeg, gif, png) représentent un peu plus de 15 % des fichiers collectés.

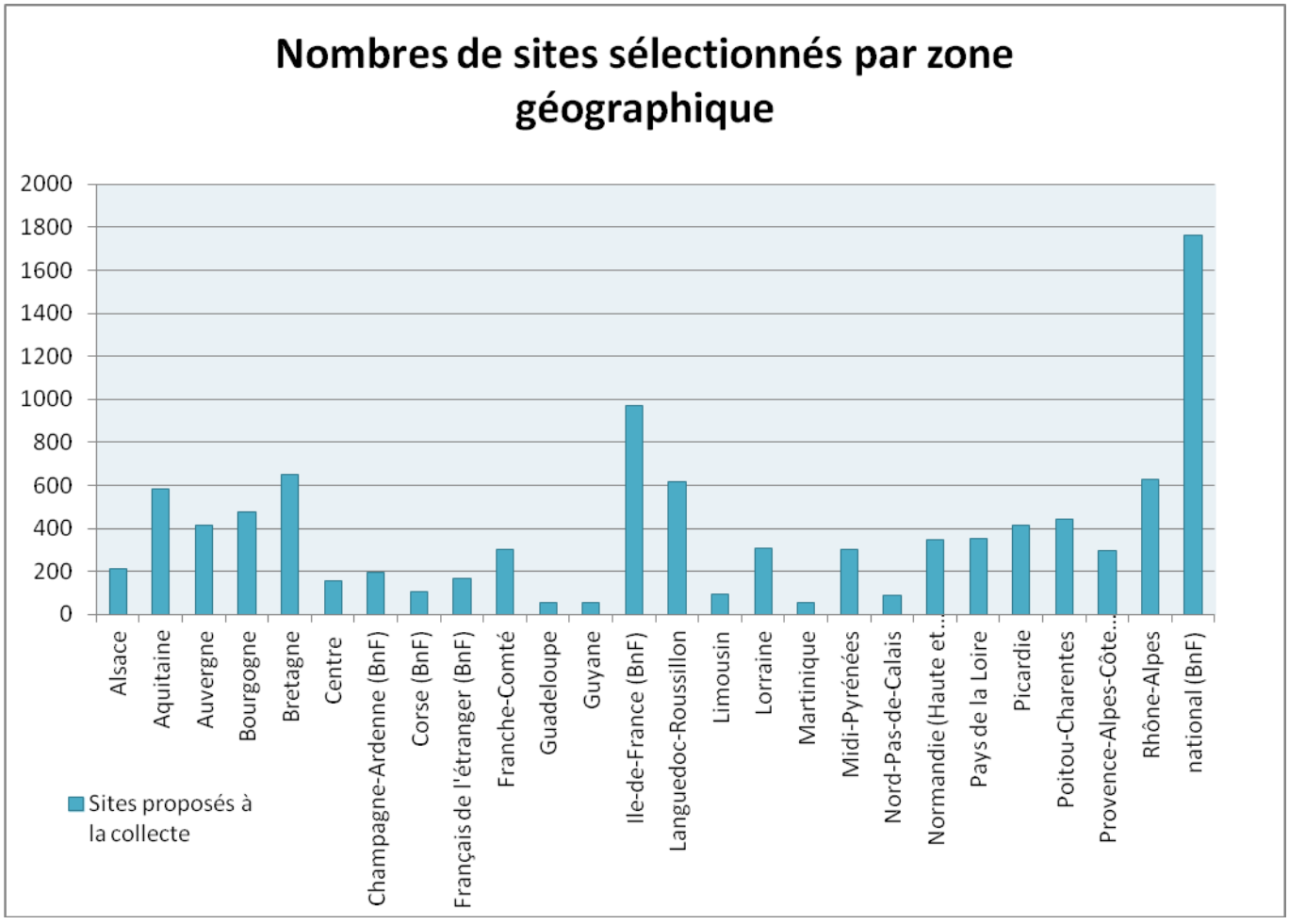
[72] Les chiffres indiqués ont été calculé à partir de l'application BnF Collecte du Web,, outil destiné à gérer des ensembles de sites Web devant être capturés. Cette application permet aux correspondants DLWeb de sélectionner et de gérer (saisir, modifier, inactiver) des sites à collecter en saisissant leur adresse URL, des paramètres techniques et une description documentaire. Ainsi, il est possible d'afficher tous les sites de la collecte Elections 2012 pour en connaître le nombre exact.

[73] Les chiffres indiqués dans ce paragraphe et dans le suivant, ont été calculés par le DLN à partir d'un outil de statistiques fournis par le département des Systèmes d'Information (les chiffres proviennent du logiciel de capture Heritrix).

2. Analyse des sites proposés à la collecte

Dans le cadre d'une réflexion sur ce que constitue un périmètre français de l'Internet, il est possible de rechercher quels sont les domaines de haut niveau les plus représentés au sein de la collection. Sans surprise, le .fr est le plus représenté, suivent les domaines de haut niveau génériques .com, .org et .net. Le .com est l'extension la plus utilisée par les plateformes de blogs. Quant au .org, il s'agit d'une extension assez répandue au niveau des sites politiques, qui veulent éviter d'afficher un suffixe commercial.

Par ailleurs, on peut également connaître le nombre de sites sélectionnés par zone géographique :



Sans surprise, le plus grand nombre de sites sélectionnés correspond au niveau national (1 800 sites). Ces sites sont dans la très grande majorité des cas liés à l'élection présidentielle.

En moyenne, les BDLI ont sélectionné 330 sites chacune. Toutefois, de fortes disparités existent en fonction de la taille des régions puisque la Guyane, la Martinique, la Guadeloupe ou encore le Limousin n'ont proposé qu'une centaine de sites. A l'inverse, la Bretagne, l'Aquitaine, Rhône-Alpes ou encore le Languedoc-Roussillon ont chacune proposé plus de 600 sites. Les régions pour lesquelles les correspondants BnF ont assuré la sélection comptent en moyenne 200 sites, sauf pour l'Île-de-France qui concentre à elle seule presque 1 000 sites sélectionnés.

La répartition par domaine de haut niveau constitue un moyen d'identifier quelles formes de publications sont représentées au sein des archives du Web électoral de 2012 à la BnF. Tandis que la répartition des sélections en fonction de la zone géographique permet de connaître la localisation des sites, et savoir dans une certaine mesure s'il s'agit de sites traitant de la présidentielle ou des législatives[74]. Pour autant, ces informations ne donnent aucune indication sur le contenu même des archives. A cette fin, il est possible d'utiliser le travail mené par les sélectionneurs lors de la saisie des sites dont ils demandaient la collecte. Parmi les différents descripteurs renseignés, le champ « type de site » (champ obligatoire) permet de déterminer la répartition des sites en fonction de la typologie.

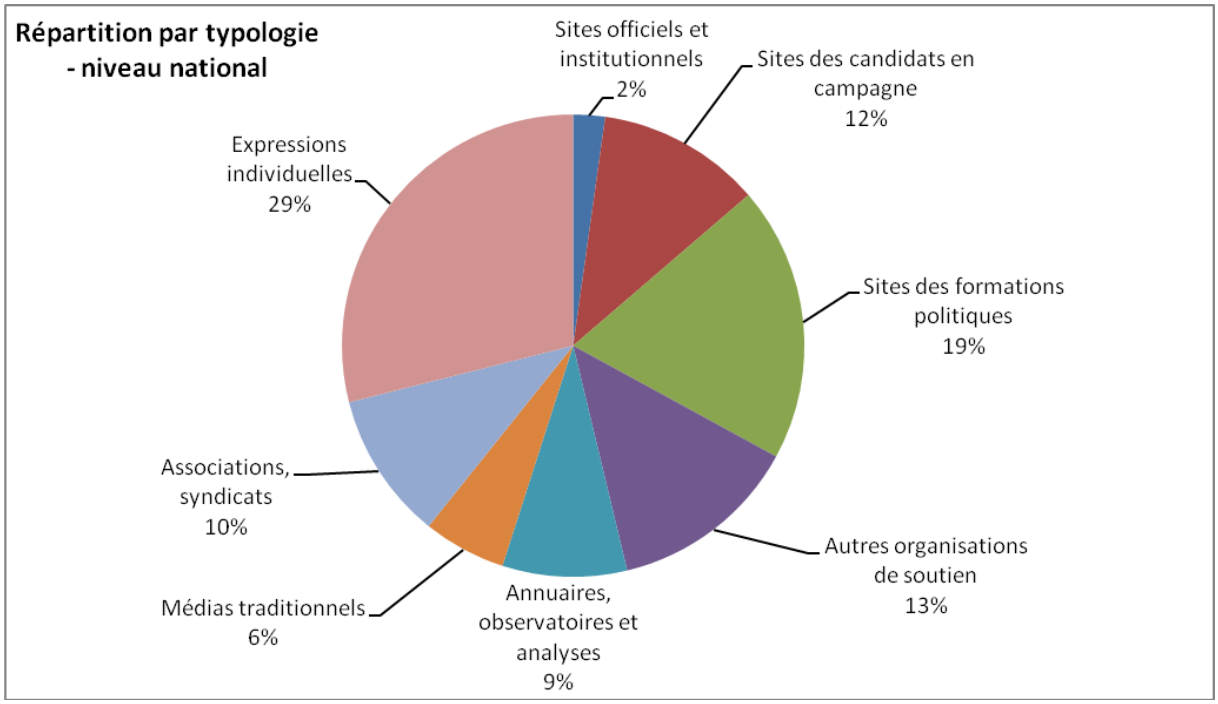
Cette répartition par type de site peut être affinée. La proportion entre types de sites est tout à fait différente lorsqu'il s'agit des sites sélectionnés au niveau national (élection présidentielle majoritairement) et au niveau régional (élection législative majoritairement).

Type de site	**Nombre de sites sélectionnés au niveau national**
0 officiels et institutionnels	39
1.1 candidats en campagne	202
1.2 formations politiques	340
1.3 autres organisations de soutien	234
2.1 annuaires, observatoires et analyses	153
2.2 médias traditionnels	102

[74] Dans la mesure où les BDLI saisissent en grande majorité des sites traitants des législatives.

2.3 associations, syndicats et autres organisations 182

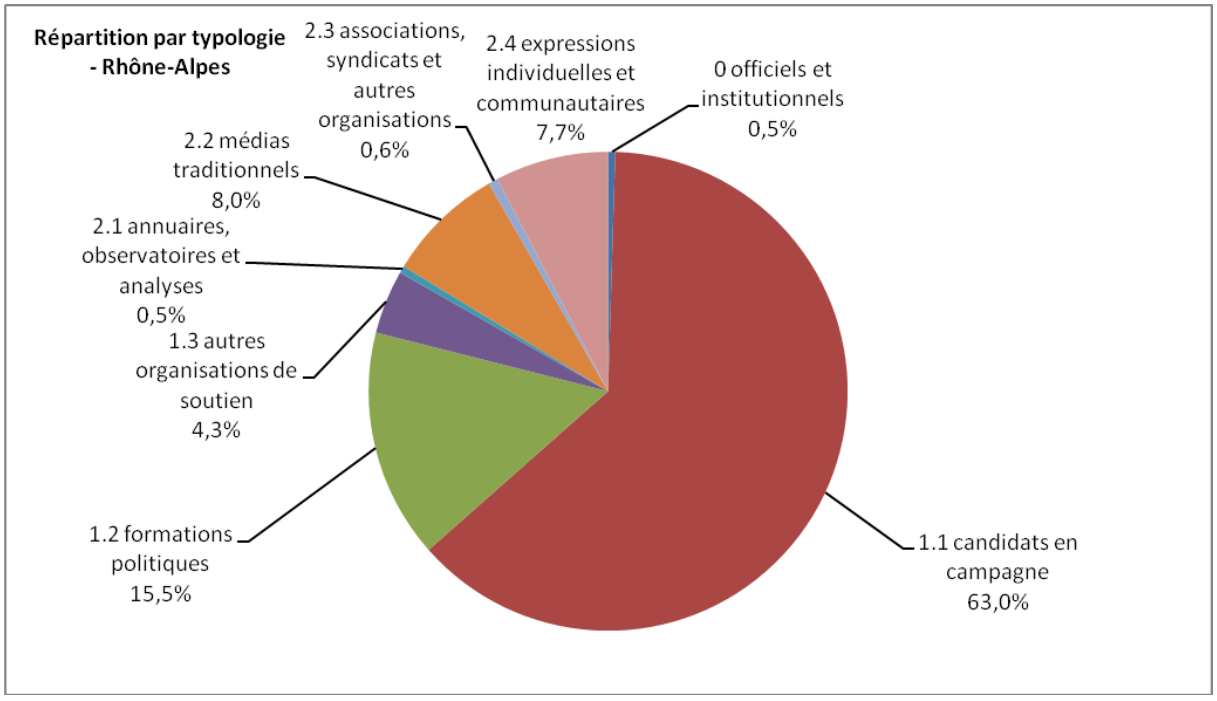
2.4 expressions individuelles et communautaires 510



Les proportions sont similaires aux scrutins de 2002 et de 2007 avec une répartition relativement équilibrée où seul le type 0 (Sites officiels et institutionnels) est peu représenté, s'expliquant simplement par le nombre peu élevé de sites de cette catégorie présents sur la Toile. Comme lors des précédents scrutins, ce sont les sites d'acteurs s'exprimant hors du cadre politique ou médiatique traditionnel qui représentent la majorité des sites saisis (30 %). Quant aux sites des partis politiques ou des candidats, ils représentent plus d'un tiers des propositions.

Si l'on compare cette répartition avec celle des sites sélectionnés à l'échelle régionale, en prenant pour exemple les sélections de la région Rhône-Alpes :

Type de site	Nombre de sites sélectionnés par la région Rhône-Alpes
0 officiels et institutionnels	3
1.1 candidats en campagne	395
1.2 formations politiques	97
1.3 autres organisations de soutien	27
2.1 annuaires, observatoires et analyses	3
2.2 médias traditionnels	50
2.3 associations, syndicats et autres organisations	4
2.4 expressions individuelles et communautaires	48



La catégorie largement dominante est désormais celle des candidats en campagne. Ceci correspond à la politique de sélection retenue par les BDLI : l'identification des sites du type 1.1 était la priorité des sélectionneurs. Les autres catégories reculent non seulement en proportion du total, mais aussi en valeur absolue. En fait, de nombreux sites de ces catégories, avaient été déjà repérés pour l'élection présidentielle au niveau national. De plus, il est difficile de repérer sur la Toile ces catégories de sites au niveau régional.

Le nombre de sites de candidats ne doit pas surprendre, dans la mesure où les élections législatives ont vu se présenter 604 candidats pour 52 circonscriptions[75] rien que pour la région Rhône-Alpes.

Toutes ces analyses permettent de mieux caractériser la collection Elections 2012. D'autres analyses sont possibles comme la répartition par formation politique, qui permet de voir quel parti est bien implanté sur le Web par exemple.

3. Les particularités de la collection électorale 2012

Comme nous l'avons indiqué précédemment, la répartition des sites au sein de la typologie n'a pas beaucoup évolué depuis 2002. Les sites d'acteurs s'exprimant hors du cadre politique ou médiatique traditionnel sont toujours les plus nombreux au niveau national, tandis que les sites de candidats en campagne sont les plus nombreux au niveau régional. Cependant, la collecte du Web électoral de 2012 a ses particularités. En effet, depuis 2002, le Web a beaucoup évolué. L'émergence des réseaux sociaux, des applications Web ou encore la présence plus marquée des contenus audiovisuels ont contribué à l'évolution des Web campagnes.

Les contenus audiovisuels représentent une part de plus en plus importante sur le Web. Pour la campagne 2012 la collecte des vidéos a constitué une part importante des sites sélectionnés, avec une implication spécifique du département Audiovisuel. La BnF a mis au point un modèle spécifique de collecte afin de capturer les vidéos de la plateforme Dailymotion[76]. Certaines vidéos ont fait l'objet de sélections à l'unité, afin de garantir leur capture. Le choix de ces vidéos pouvait reposer sur des critères variés : popularité (nombre de visiteurs), représentativité, spécificité du message par rapport aux autres médias… La plateforme Dailymotion a également permis une sélection par le compte des candidats, ce qui offrait l'avantage de faire collecter ensuite par le robot la totalité des vidéos postées sur un compte.

Du côté des réseaux sociaux, en 2012, on constate une utilisation massive de ces réseaux par les candidats, notamment Facebook et Twitter[77]. Cet état de fait se ressent également dans la

[75] Source : http://www.elections-legislatives.fr, consulté le 11 septembre 2012

[76] L'architecture de ces plateformes étant très mouvante, les institutions chargées de leur capture doivent sans cesse mettre à jour leur modèle de collecte.

[77] Aussi bien Nicolas Sarkozy que François Hollande affichaient sur leur site respectif des liens vers leurs comptes Facebook et Twitter. François Hollande proposait de suivre les coulisses de sa campagne sur Tumblr tandis que Nicolas Sarkozy invitait les internautes à suivre tous ses déplacements sur Foursquare.

collection : plus d'une centaine de pages Facebook ont été collectées quotidiennement[78], ainsi qu'une centaine de comptes Twitter qui ont pour leur part été collectés quatre fois par jour. Seules les pages publiques (que l'auteur configure pour être accessibles à tous) peuvent être capturées par le robot. Ceci est très souvent le cas pour des pages à contenu politique, puisque l'auteur veut être lu par le plus grand nombre. Notons cependant que dans les archives constituées à partir de la Toile politique française, le modèle du blog conserve une place importante. De plus, cette année, la stratégie des partis politiques a évolué vers la constitution de plateformes d'hébergement qui leur sont propres. On peut citer l'exemple de la plateforme du Front de Gauche qui héberge les blogs tenus par ses candidats[79].

[78] La fréquence de renouvellement de ces pages étant relativement élevée, les pages étaient capturées une fois par jour.

[79] Exemple de blog : http://guillaume-quercy.candidats.frontdegauche.fr/, consulté le 14 septembre 2012

III. Les missions et les activités du stage

Le service du dépôt légal numérique (DLN), chargé du pilotage technique et organisationnel du projet, m'a recruté en tant que stagiaire. Faire ce stage à la BnF m'a permis de bénéficier des formations proposées au personnel et de visiter des services très variés au sein de la DCO[80] et de la DSR[81]. Même si la complexité et le caractère hiérarchisé de l'établissement exigent un temps d'adaptation, le fait d'être dans un petit service[82] facilite l'intégration.

Les principales activités du stage étaient :

- d'effectuer une veille active dans le domaine des technologies de l'information et des sciences politiques ;
- de suivre l'activité de sélection documentaire des agents participant au projet, au sein de la BnF ou des bibliothèques de dépôt légal imprimeur ;
- de piloter et contrôler la qualité des collectes effectuées par le robot d'archivage ;
- de valoriser le projet en cours et les collections constituées, auprès des chercheurs et auprès des partenaires nationaux et internationaux de la BnF.

A. Suivi de l'activité de sélection de sites des agents BnF et des agents des bibliothèques de dépôt légal imprimeur (BDLI)

1. Travail préparatoire et suivi du projet

Au cours de cette phase préparatoire, un travail de réflexion a été mené : définition du projet, estimation des besoins, validation de la hiérarchie, réunion des moyens nécessaires. Certains éléments, comme les critères de sélection, la typologie, l'utilisation des mots-clés ou encore le fonctionnement du comité de sélection, sont restés stables par rapport aux collectes précédentes du Web politique. Toutefois, des ajustements ont vu le jour au cours du projet.

[80] Visite des services des départements Droit, économie, politique et Philosophie, histoire, sciences de l'homme ; ces deux départements ont participé à la collecte du Web électoral 2012.

[81] Visite du service Support et production (département des Systèmes d'information).

[82] Le service du DLN est composé de cinq agents, dont le chef de service.

Les choix d'ordre organisationnel comme la constitution d'une équipe de sélectionneurs large issue de la BnF et des BDLI, sont également restés inchangés.

L'équipe des sélectionneurs avait été constituée dès décembre 2011. Le projet a regroupé un total de dix-huit sélectionneurs de la BnF[83]. La sélection a été répartie selon les départements soit par catégories de la typologie (Philosophie, histoire, sciences de l'homme), soit par partis politiques et candidats (Droit, économie, Politique). Pour les élections législatives, le principe géographique est devenu le critère de répartition de référence. L'ensemble du territoire (hormis les zones assurées par les BDLI) était couvert.

Le calendrier du projet a été défini ainsi :

23 janvier 2012	Lancement de la première collecte Élections mensuelle
31 janvier	Première réunion du groupe Élections BnF
1er février	Lancement des premières collectes Elections quotidienne et Elections hebdomadaire
1er mars	Lancement de la première collecte Elections quatre fois par jour (Twitter)
8 mars	Deuxième réunion de suivi du groupe Élections BnF
29 mars	Rencontre avec les BDLI impliquées
30 mars	Réunion du groupe Élections BnF : répartition des régions non couvertes par les BDLI participantes
22 avril	Premier tour de l'élection présidentielle
24 avril	Réunion d'entre-deux-tours du groupe Élections
6 mai	Second tour de l'élection présidentielle
22 mai	Réunion d'entre-scrutins du groupe Élections
10 juin	Premier tour des élections législatives
17 juin	Second tour des élections législatives
19 juin	Dernières collectes Elections quotidienne et Elections quatre fois par jour
26 juin	Réunion de bilan du groupe Élections
27 juin	Dernière collecte Elections hebdomadaire
29 juin	Fin des saisies dans BCWeb
24 juillet	Dernière collecte Élections mensuelle
À partir de juin-juillet	Bilan et valorisation

[83] Neuf sélectionneurs du Département Droit, économie, politique ; sept du Département Philosophie, histoire, sciences de l'homme et deux du Département Audiovisuel. Il faut ajouter à ces chiffres un à quatre sélectionneurs par BDLI participante au projet, soit un total d'environ soixante-dix personnes.

Un comité de sélection a été mis en place, constitué de l'ensemble des sélectionneurs, ainsi que des pilotes documentaires et techniques. Comme indiqué sur le calendrier, il s'est réuni, au cours du projet à six reprises, le rythme de réunion dépendant largement de l'actualité (le rythme était plus resserré autour des scrutins). Ce comité de sélection qui était donc le principal organe opérationnel du projet, avait deux rôles majeurs, un rôle de diffusion des informations et un rôle de concertation.

- Le rôle de diffusion : chaque réunion commence par un point d'actualité, concernant en général le Web politique (événements de la campagne, polémiques, nouveauté dans les médias…). Au cours de chaque réunion, un point d'avancement était fait, pour indiquer le rythme de progression du nombre de propositions. Les collectes en cours et à venir étaient également précisées aux sélectionneurs. Par ailleurs, les problèmes techniques rencontrés par les sélectionneurs au cours des dernières semaines étaient abordés et résolus.
- Le rôle de concertation : ces réunions étaient également le lieu pour débattre de nombreuses questions : fallait-il sans cesse modifier mots-clés pour coller à l'actualité (nom des partis politiques, candidats dissidents…) ? Quelle attitude adopter vis-à-vis des sites de presse en ligne (qui constituaient souvent un doublon avec les autres collectes menées par la BnF[84]) ? Fallait-il saisir les sites d'hommes politiques d'envergure nationale, mais qui ne parlaient pas de la campagne électorale : leur silence à ce sujet pouvait-il être considéré comme signifiant, donc digne d'être collecté ? (exemple du site de Christiane Taubira, qui n'était pas candidate aux législatives)

Ces réunions du comité de sélection ont fait l'objet de comptes rendus détaillés ou bien de messages envoyés au groupe Élections. En effet, il était nécessaire de garder une mémoire des actions menées et des décisions prises.

Quant aux tâches effectuées par l'équipe du Dépôt légal du Web (DLWeb), on peut reprendre la représentation utilisée pour représenter le processus de dépôt légal de la Toile[85]. Ce terme d'équipe permet de regrouper les personnes qui, ne travaillant pas nécessairement dans le même département, sont affectées à plein temps sur le projet de DLWeb. Il s'agit, au sein du

[84] A côté des collectes thématiques de chaque département de la BnF, il existe d'autres collectes courantes. On peut citer celle de la presse ou encore celle consacrée aux publications officielles.

[85] Voir en annexe le Schéma directeur du processus de dépôt légal de la Toile à la BnF (document BnF), ce dernier expose en détail chaque étape.

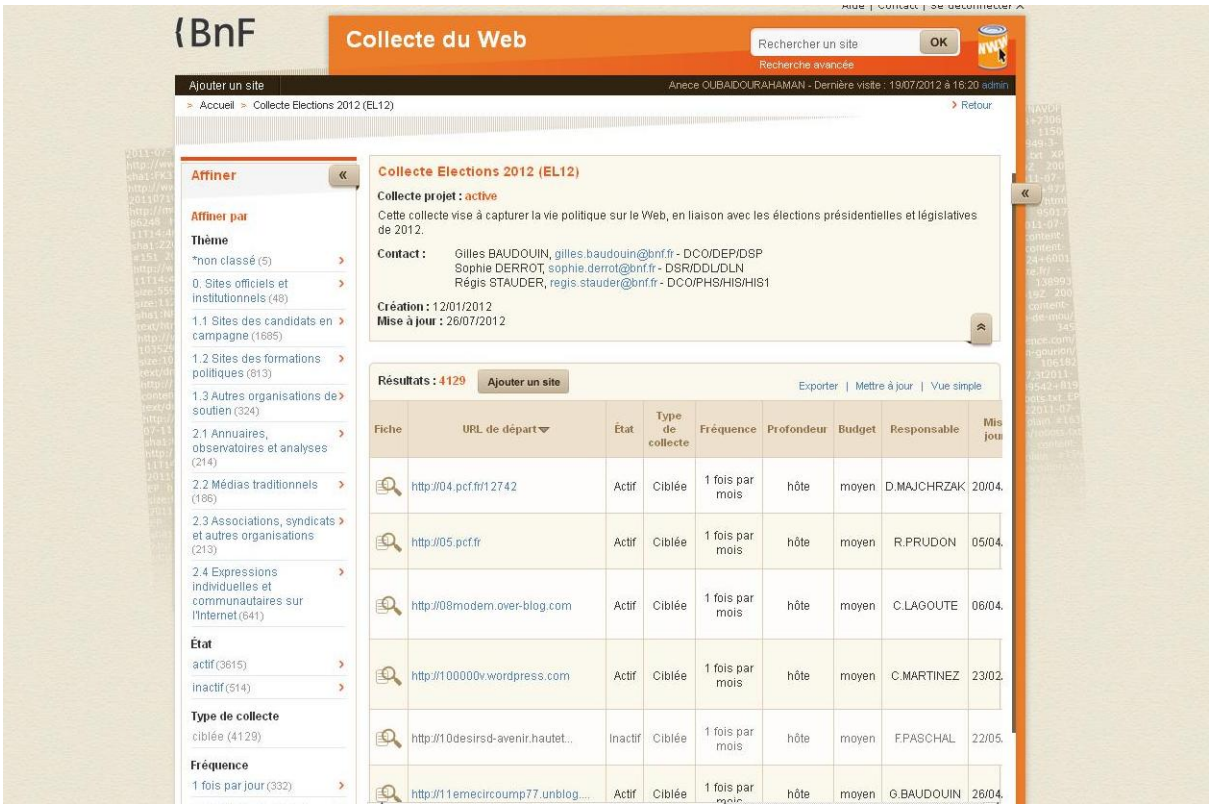
département de Dépôt légal, du service du Dépôt légal numérique (quatre agents et un stagiaire) et de deux agents du département des Systèmes d'information.

2. Vérification des propositions d'URL à collecter

Durant toute la durée du projet, les sélectionneurs saisissent des propositions d'URL à collecter sur l'application BnF Collecte du Web[86] (BCWeb), au fur et à mesure qu'ils découvrent les sites. L'application BCWeb est un outil destiné à gérer des ensembles de sites devant être capturés. L'outil permet donc de saisir les adresses des sites à collecter, à leur attribuer un certain nombre de paramètres techniques et de descriptions d'ordre documentaire. Il permet aussi de gérer les sites, les paramètres et les descriptions saisies. L'application fournit à dates régulières (définies par le DLWeb) des listes de sites aux robots de collecte. BCWeb organise les propositions des correspondants sous formes de fiches de site, qui sont regroupées dans des collectes. Il est important de préciser que BCWeb ne regroupe pas les références de tous les sites Internet archivés par la BnF au titre du dépôt légal, puisque comme nous l'avons vu précédemment, la Bibliothèque réalise également des collectes larges de l'ensemble de l'Internet français. Au vu du nombre de sites (plus de deux millions en 2011) et étant donné qu'ils ne sont pas identifiés par avance, ces sites ne sont pas intégrés dans BCWeb.

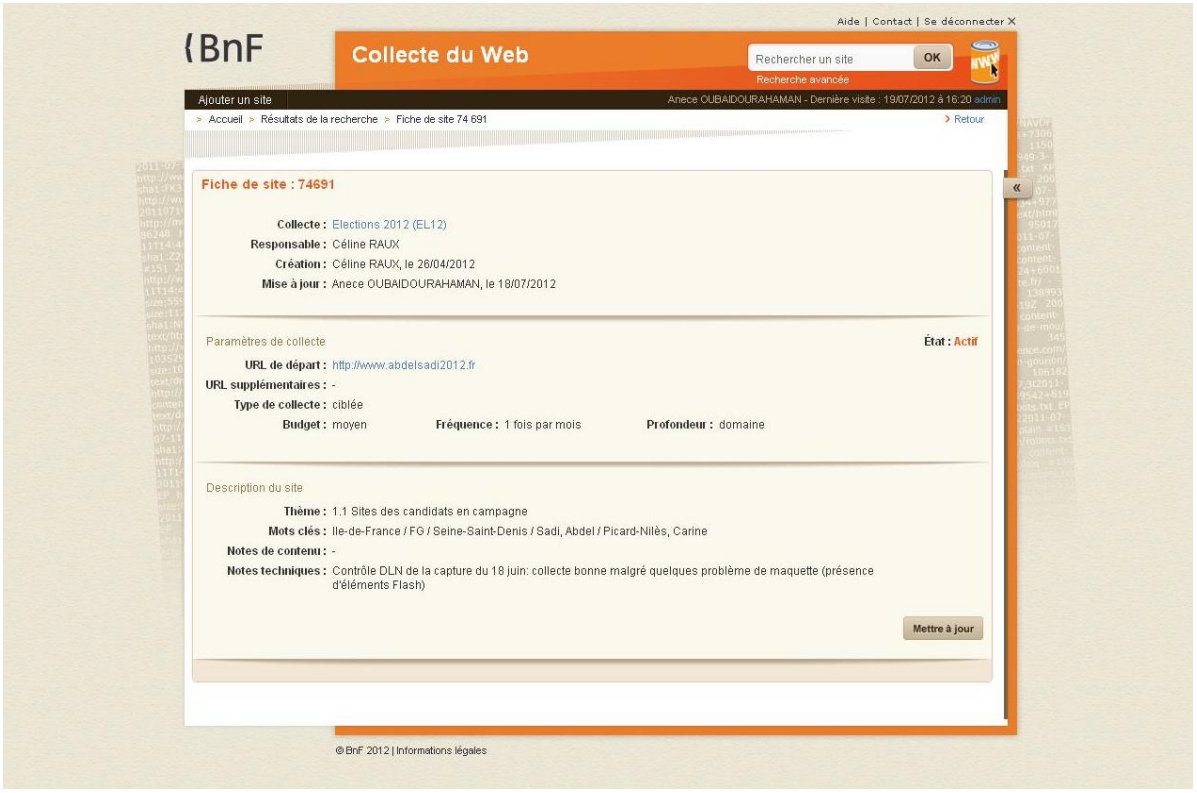
[86] Accessible en ligne par l'adresse suivante, uniquement sur identification : https://collecteweb.bnf.fr

La collecte Elections 2012 se présente ainsi dans BCWeb :



- Le cadre supérieur donne une description rapide du projet, les contacts des coordonnateurs du projet (un responsable par établissement chargé de la sélection et un responsable du service de Dépôt légal numérique) ;
- Le cadre latéral gauche offre des possibilités de tri. L'affinage s'effectue en fonction de critères documentaires (typologie) ou de critères techniques (paramètres de collecte).

Un clic sur l'icône représentant une loupe, permet d'accéder à la fiche d'un site :



Une fiche de site se décompose en 3 parties que nous verrons en détail par la suite :

- en haut, la partie de gestion de l'application : collecte concernée, responsable, créateur et dernier modificateur de la fiche ;
- au centre, les paramètres techniques choisis pour la collecte ;
- en bas, la description documentaire de la proposition.

Chaque agent du DLN a à sa charge la vérification des collectes dont il est responsable[87]. La vérification des URL consiste à contrôler que les sites proposés sont techniquement capturables (des animations Flash, des formulaires de requête ou encore la présence de # dans les URL peuvent en effet constituer un obstacle pour le robot), que les URL de départ spécifiées sont les plus pertinentes et que les paramètres de fréquence et de profondeur choisis sont cohérents par rapport au contenu du site

Concernant la vérification des paramètres de collecte, il s'agit de voir si le budget, la fréquence et la profondeur choisis conviennent à l'URL proposée. Lors de l'ajout d'une nouvelle proposition, dans chacun des menus déroulants s'affichent les valeurs autorisées pour la collecte Élections (qui peuvent donc différer des autres collectes).

[87] Sur la collecte Elections 2012, chaque agent du DLN devait traiter cinq à six collectes, chaque BDLI disposant de sa propre collecte électorale.

Le budget indique le nombre d'URL à collecter sur le site proposé. Il est donc attribué en fonction de l'estimation de la taille d'un site[88]. Pour la collecte Élections, les sélectionneurs avaient le choix entre moyen (150 000 URL) et micro (10 000 URL).

La fréquence indique la périodicité de la collecte. Ce paramètre est défini en fonction du taux de renouvellement d'un site ou de la partie visée d'un site. Pour le projet Élections, les fréquences possibles étaient : pluriquotidienne (quatre fois par jour)[89], quotidienne, hebdomadaire, mensuelle et après scrutins.

La profondeur délimite le contour de la collecte. Elle indique si la collecte porte sur la totalité du site ou sur une partie. Pour la collecte Elections, les profondeurs proposées étaient les suivantes : domaine, hôte, chemin, page +2, page+1 actu et vidéo. Dans le cadre d'une collecte projet, ce paramètre est peut-être le plus important. En effet, une profondeur trop restreinte ne permettrait pas de collecter suffisamment, tandis qu'une profondeur trop large apporterait trop de bruit et surchargerait les serveurs[90] :

- Domaine : profondeur de collecte qui correspond à la totalité d'un site (exemple : http://eelv.fr). Cette profondeur a souvent été utilisée pour capturer le site d'un candidat.
- Hôte : profondeur qui permet de collecter une certaine partie du site hébergée de manière individuelle (exemple : http://creteil.eelv.fr). Cette profondeur est adaptée pour la sélection des blogs hébergés sur une plateforme (free.fr, hautetfort.com, lesdemocrates.fr).
- Chemin : profondeur qui permet de collecter tous les fichiers contenus dans une rubrique d'un site, ou dans un répertoire, chaque répertoire s'ouvrant à partir du domaine principal avec un « / ». (exemple : http://creteil.eelv.fr/nos-elues/)
- Page +2 : profondeur qui permet de collecter une page simple ainsi que les contenus se trouvant à 2 clics de cette page et sur le même domaine. Cette profondeur est notamment pratique pour capturer les pages Wikipédia.

[88] Pour estimer la taille d'un site faire une recherche sur Google avec « site: » (exemple : site:lesdemocrates.fr donne un résultat de 160 000 pages).

[89] Concerne uniquement la collecte de Twitter.

[90] La définition des paramètres de fréquence est indiquée dans le document BnF intitulé « BCWeb : manuel de l'utilisateur » et destiné aux sélectionneurs.

- Page+1 : profondeur qui correspond à la celle de la collecte Actualités[91] et permet de relever rapidement et efficacement des contenus se renouvelant très fréquemment. Cette profondeur permet donc de collecter la page en question plus tous les fichiers se trouvant à un clic de celle-ci. La profondeur page +1 a notamment été utilisée pour capturer les réseaux sociaux qui font preuve d'une activité dense et rapide. (exemple : http://www.facebook.com/bayrou)
- Vidéo : cette profondeur était spécifiquement destinée à la collecte des vidéos Dailymotion.

Ainsi, pour le robot de collecte la page est l'unité la plus petite et la plus rapide à collecter, et le domaine la plus large, donc potentiellement la plus longue à collecter si le site contient beaucoup de données.

Le service du DLN doit donc vérifier si les paramètres techniques que nous venons de détailler ont été correctement saisis et les modifier si nécessaire. La typologie et les mots-clés descriptifs d'un site peuvent également nécessiter des modifications. Dans ce cas, l'agent du DLN prend contact avec le sélectionneur du site en question pour en discuter avec lui. D'une manière générale, les corrections apportées par le DLN font toujours l'objet d'une communication vers les sélectionneurs, par l'intermédiaire d'échange de courriel ou bien lors de la réunion du comité de sélection. De plus, une zone de notes sur la fiche de site dans BCWeb permet d'indiquer les modifications effectuées.

Dans le cadre d'une collecte projet, les agents du DLN peuvent également apporter leur aide afin de repérer les sites les plus pertinents. Toutefois, la décision de sélectionner ou non un site relève entièrement de la responsabilité des sélectionneurs de la Direction des Collections et des BDLI.

3. Aide à la veille et au repérage de sites

Par sa connaissance de l'Internet, le service du Dépôt légal numérique fournit des méthodes aux sélectionneurs pour trouver les sites les plus pertinents. De nombreux outils accessibles en ligne permettent aux sélectionneurs d'effectuer leur veille. C'est le cas des principaux moteurs de recherche (google.fr, yahoo.fr, bing.fr,…) qui donnent des résultats rapides et souvent intéressants. Ainsi, entrer le nom d'une région suivi de « élections 2012 » par exemple, ou bien encore celui d'une personnalité permet de découvrir les sites officiels, mais

[91] Une collecte ciblée dédiée spécifiquement à l'actualité capture quotidiennement une centaine de sites de quotidiens nationaux, régionaux et de *pure players* (sites Web d'information purement numériques). Cette collecte complète donc également la collecte Election pour la période de la campagne.

également les blogs, les sites de détracteurs, les pages de réseaux sociaux… Plus spécialisés, les annuaires et portails thématiques permettent souvent d'accéder directement à l'information sous son angle politique[92]. Quant aux annuaires de partis[93], ils permettent une veille systématique afin d'identifier tous les sites d'un parti.

Les réseaux sociaux, par leur rapidité à faire circuler l'information, apportent également des résultats appréciables (Facebook, Twitter, Google+…). Des sites Web peuvent également être consultés, tel que Partie2Campagne[94] qui offre pour chaque candidat à la présidentielle 2012 une cartographie des usages et des volumes de mots-clés de la campagne, avec un accès aux documents sources sur le Web et aux sites Internet qui les diffusent. Il existe également plusieurs Pearltrees sur les élections[95].

Pour effectuer une veille efficace, il ne faut pas non plus négliger les autres médias (radio, télévision, presse écrite…). Par ailleurs, la plupart des sites d'information en ligne créent à l'occasion des élections des rubriques spécifiques où les événements de l'actualité politique trouvent un écho intéressant.

Afin de repérer les sites, il est important d'être réactif par rapport à l'apparition ou au contraire à la disparition de sites, ainsi qu'à l'évolution majeure des contenus d'un site[96].

Une fois que le site est saisi par le sélectionneur et qu'il a été vérifié par un agent du DLN, le site part en collecte.

[92] Il s'agit des sites comme http://www.index-politique.fr/, qui couvre la vie politique française sur l'Internet au sens large, ou bien http://www.elus20.fr/services/deputes-legislatives-2012/, qui traite plus spécifiquement des élections législatives.

[93] Tel que https://www.federationsump.org/, http://europe-ecologie.net/ ou encore http://data.parti-socialiste.fr/

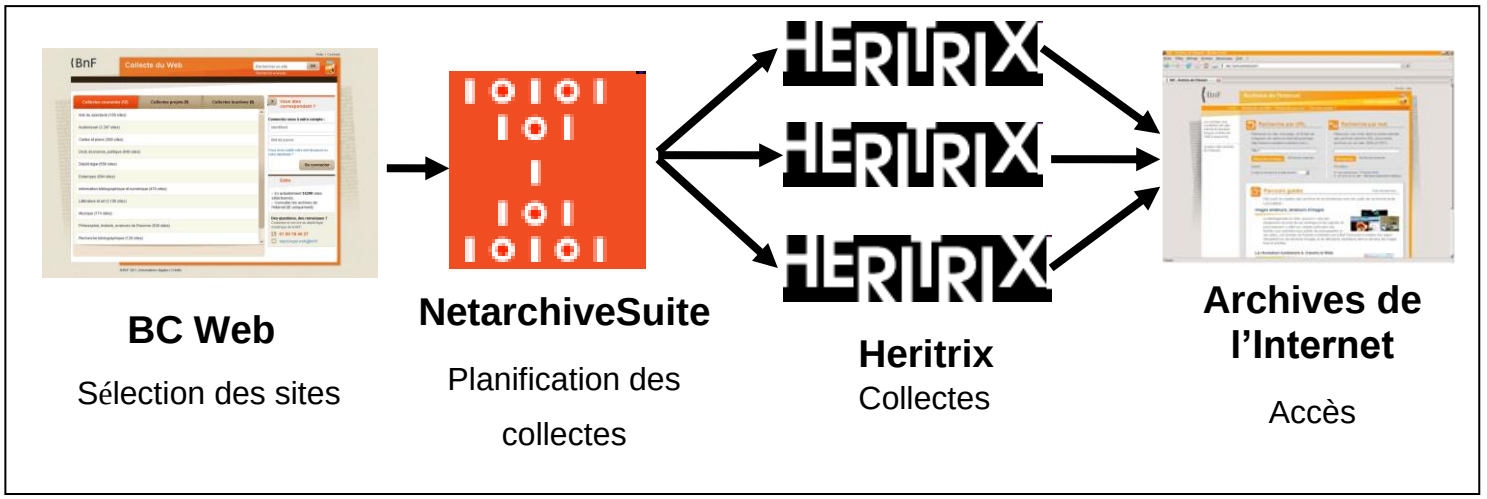
[94] http://www.partie2campagne.fr/

[95] Un Pearltree est une organisation de pages web effectué par une personne ou par un groupe de personnes. Sur les élections, on peut citer : http://www.pearltrees.com/t/election-presidentielle-france/id3955156, consulté le 12 septembre 2012.

[96] Concernant l'évolution du contenu, le site officiel du candidat Nicolas Sarkozy a marqué une pause dans la campagne électorale en diffusant un message de compassion au lendemain des tueries de Toulouse. Le reste des contenus du site n'étaient pas disponibles le temps du deuil et de l'hommage aux victimes.

B. Pilotage et contrôle de la qualité des collectes effectués par le robot d'archivage[97]

Avant d'aborder le pilotage et le contrôle qualité des collectes, il est nécessaire d'apporter quelques précisions sur la chaîne de production du dépôt légal du Web. En effet, depuis la mise en place de BCWeb début 2012, on peut parler de réel flux de production pour qualifier la collecte du Web à la BnF, puisque les données migrent automatiquement d'une application à l'autre :



1. Lancement d'une collecte

Le lancement d'une nouvelle collecte s'effectue suivant le planning de suivi technique et fonctionnel 2012. Ce dernier indique les dates de collecte des « ensembles à collecter ». Un ensemble à collecter comprend tous les sites d'une même collecte ayant des paramètres de fréquence et de profondeur identiques. Le DLN s'est doté d'une structure normalisée pour établir le nom des ensembles à collecter, construite ainsi : BnF nom de la collecte fréquence budget.

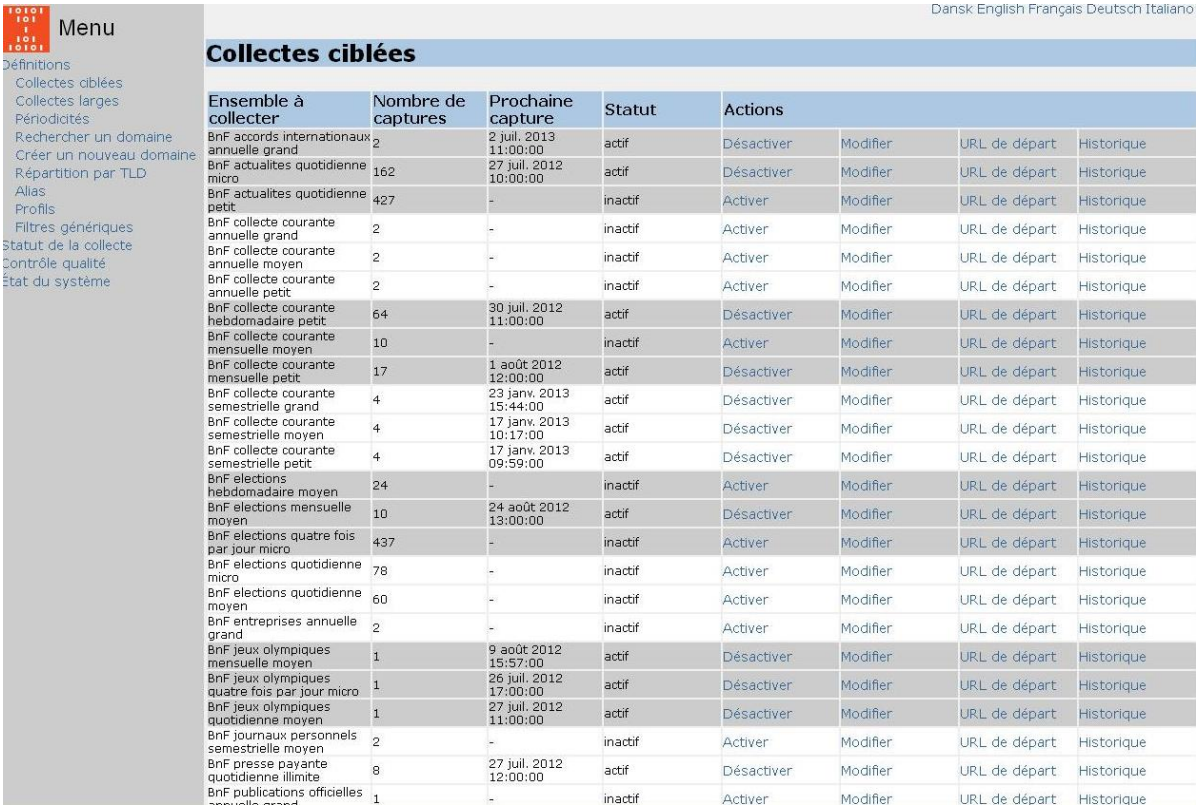
Le planning de suivi technique et fonctionnel est régulièrement revu lors des réunions hebdomadaires entre le DLN et les ingénieurs de production[98].

[97] Les informations qui suivent sont tirées du document BnF intitulé « Memento de production : formalisation des opérations internes au DLN, année 2012 ».

[98] Il s'agit d'une réunion hebdomadaire entre le DLN et les ingénieurs du Service Support et production (DSR) où un point est fait sur les collectes en cours. Les collectes à venir sont également discutées d'un point de vue technique.

A l'approche d'une collecte planifiée, le responsable DLN prépare le transfert de la collecte en question vers NetarchiveSuite (NAS)[99]. La préparation consiste à vérifier la syntaxe des URL saisies profondeur par profondeur, à rendre inactif les sites saisies mais relevant du périmètre de l'INA, et de relever le nombre de configurations afin de contrôler par la suite le bon déroulement du transfert. Une fois le transfert effectué, une dernière vérification des listes d'URL de départ est nécessaire.

L'application NAS se présente sous cette forme et permet donc de sélectionner les listes d'URL de départ en fonction des collectes :



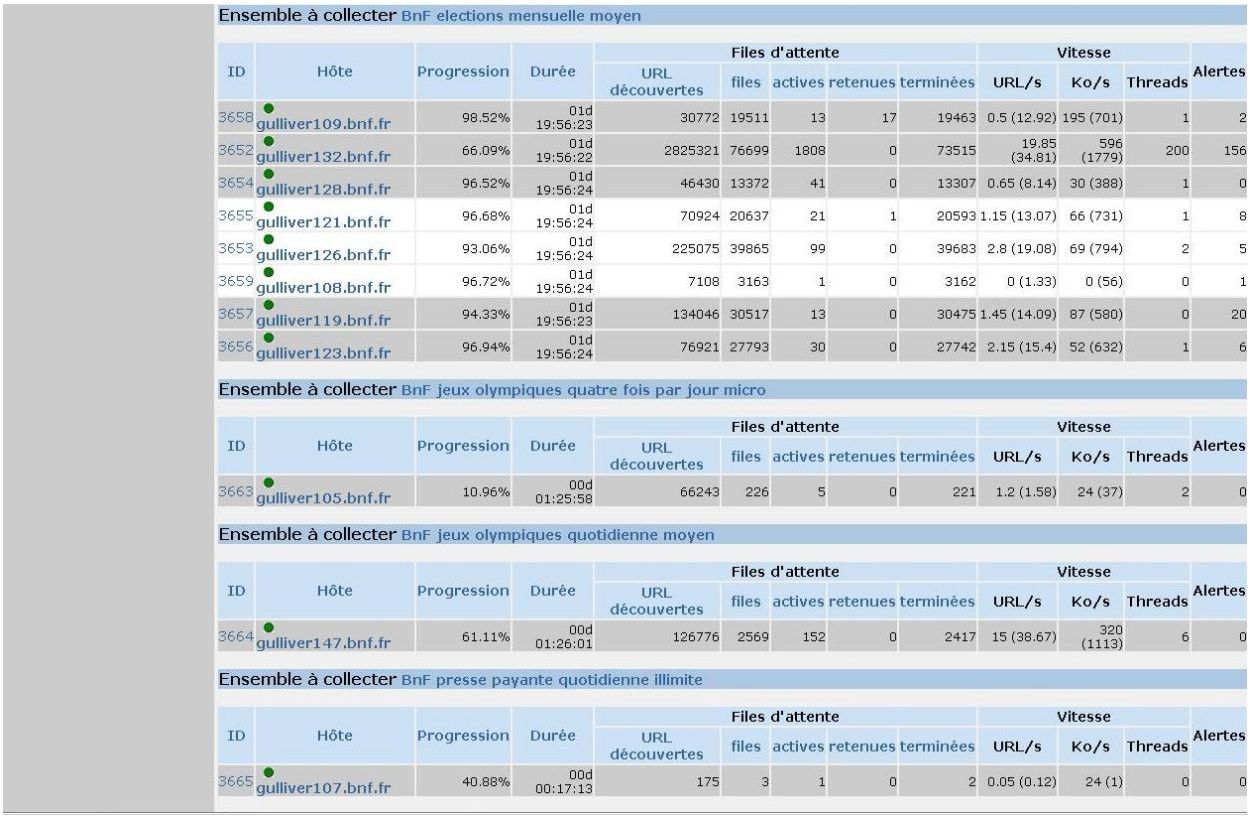
2. Surveillance des collectes en cours

Durant la phase de collecte, la surveillance est une des activités les plus importantes effectuées par un agent du DLN. Le logiciel NAS partage les ensembles à collecter en plusieurs paquets appelés « jobs » en fonction de la profondeur de collecte. Cette division facilite la vérification durant la collecte et permet au DSI de répartir les collectes sur les divers serveurs dédiés au moissonnage. Chaque agent du DLN a un certain nombre de jobs à

[99] Le transfert pour les collectes courantes aux fréquences quotidiennes, hebdomadaires et mensuelles est réalisé à la fin de chaque mois. Pour les autres collectes courantes et projets, le transfert est réalisé juste avant la date de lancement de la collecte.

surveiller. La surveillance se fait d'une part grâce à l'application NAS, et d'autre part par l'intermédiaire du logiciel de collecte Heritrix.

NAS permet de visualiser les collectes en cours :



Ensemble à collecter BnF elections mensuelle moyen

ID	Hôte	Progression	Durée	Files d'attente					Vitesse			Alertes
				URL découvertes	files	actives	retenues	terminées	URL/s	Ko/s	Threads	
3658	gulliver109.bnf.fr	98.52%	01d 19:56:23	30772	19511	13	17	19463	0.5 (12.92)	195 (701)	1	2
3652	gulliver132.bnf.fr	66.09%	01d 19:56:22	2825321	76699	1808	0	73515	19.85 (34.81)	596 (1779)	200	156
3654	gulliver128.bnf.fr	96.52%	01d 19:56:24	46430	13372	41	0	13307	0.65 (8.14)	30 (388)	1	0
3655	gulliver121.bnf.fr	96.68%	01d 19:56:24	70924	20637	21	1	20593	1.15 (13.07)	66 (731)	1	8
3653	gulliver126.bnf.fr	93.06%	01d 19:56:24	225075	39865	99	0	39683	2.8 (19.08)	69 (794)	2	5
3659	gulliver108.bnf.fr	96.72%	01d 19:56:24	7108	3163	1	0	3162	0 (1.33)	0 (56)	0	1
3657	gulliver119.bnf.fr	94.33%	01d 19:56:23	134046	30517	13	0	30475	1.45 (14.09)	87 (580)	0	20
3656	gulliver123.bnf.fr	96.94%	01d 19:56:24	76921	27793	30	0	27742	2.15 (15.4)	52 (632)	1	6

Ensemble à collecter BnF jeux olympiques quatre fois par jour micro

ID	Hôte	Progression	Durée	Files d'attente					Vitesse			Alertes
				URL découvertes	files	actives	retenues	terminées	URL/s	Ko/s	Threads	
3663	gulliver105.bnf.fr	10.96%	00d 01:25:58	66243	226	5	0	221	1.2 (1.58)	24 (37)	2	0

Ensemble à collecter BnF jeux olympiques quotidienne moyen

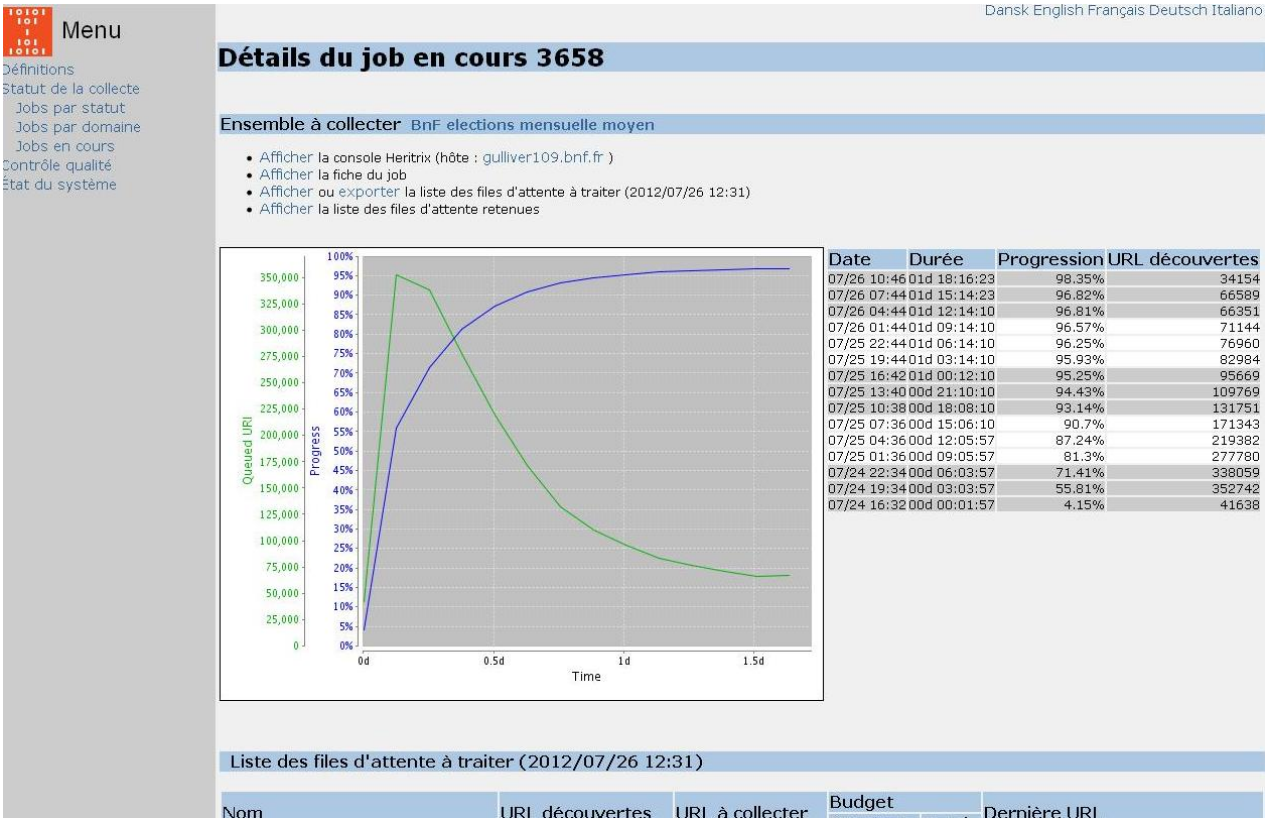
ID	Hôte	Progression	Durée	Files d'attente					Vitesse			Alertes
				URL découvertes	files	actives	retenues	terminées	URL/s	Ko/s	Threads	
3664	gulliver147.bnf.fr	61.11%	00d 01:26:01	126776	2569	152	0	2417	15 (38.67)	320 (1113)	6	0

Ensemble à collecter BnF presse payante quotidienne illimite

ID	Hôte	Progression	Durée	Files d'attente					Vitesse			Alertes
				URL découvertes	files	actives	retenues	terminées	URL/s	Ko/s	Threads	
3665	gulliver107.bnf.fr	40.88%	00d 00:17:13	175	3	1	0	2	0.05 (0.12)	24 (1)	0	0

Lorsqu'un job échoue, le DLN fait une évaluation de la pertinence de relancer ou non le job grâce aux précisions techniques apportées par les ingénieurs de production.

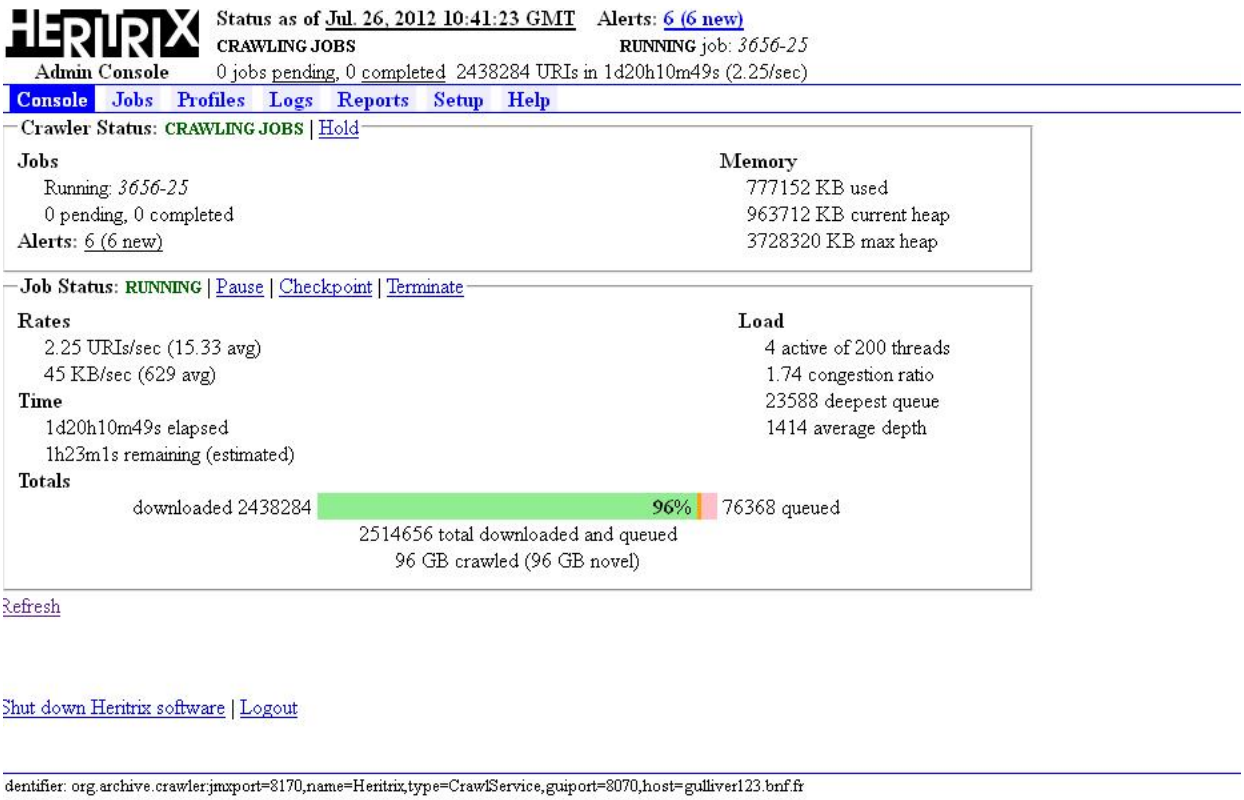
Pour surveiller un job, il faut accéder aux détails d'un job en cours :



A partir de cet écran, il est possible :

- d'analyser les courbes de progression des jobs,
- de vérifier en ligne les files d'attentes à traiter et d'identifier les boucles,
- d'analyser les files retenues et éventuellement les relancer.

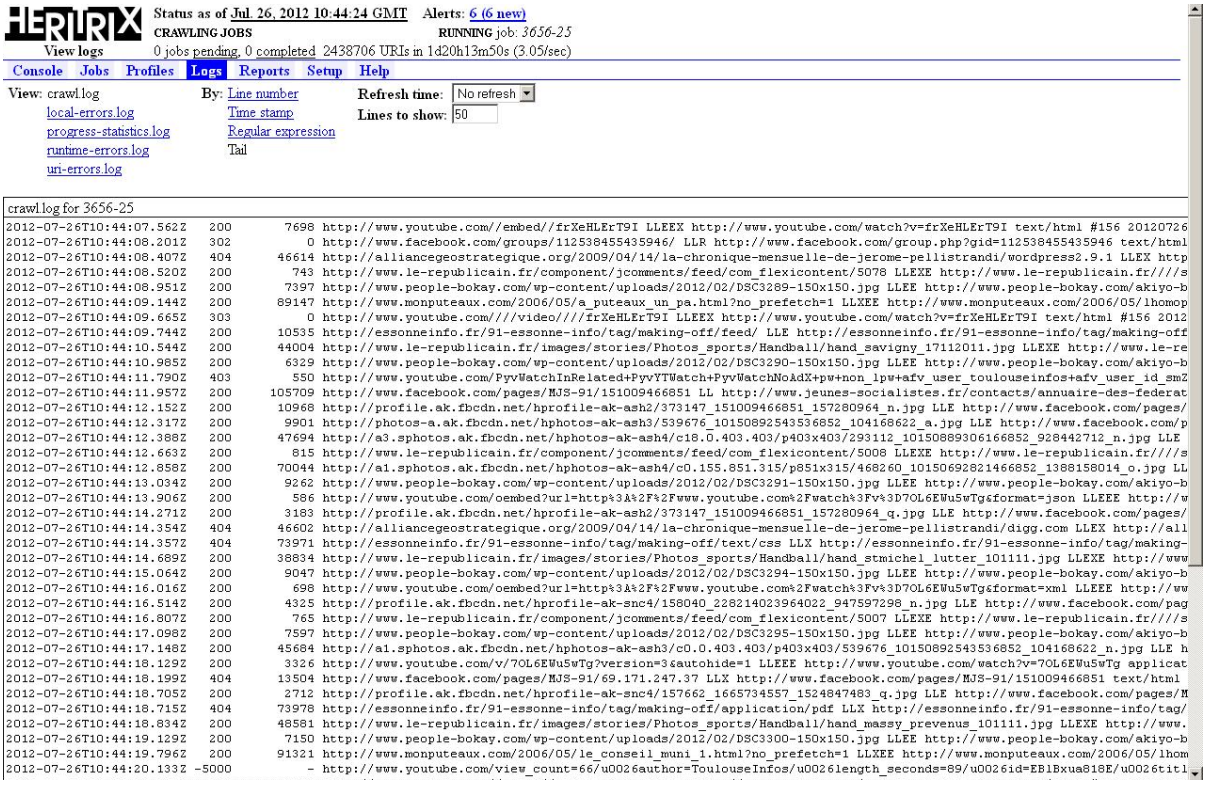
NAS permet d'accéder également au serveur de collecte Heritrix qui concerne la collecte surveillée, le serveur Heritrix se présente ainsi :



Sur un serveur Heritrix, il est possible :

- de vérifier les redirections et les reporter dans BCWeb ;
- de vérifier ce qui est en train d'être collecté depuis la page des « logs », ;
- d'étudier un hôte particulier depuis le « crawl report » ;
- de filtrer les URL indésirables.

La page des « logs » se présente ainsi, et permet de voir en direct les données qui entrent dans la collection :



Lors de cette phase de collecte, il est possible que des producteurs de site mécontents (surcharge de leurs serveurs provoquée par la collecte par exemple) s'adressent au DLN. La réactivité du DLN sur la boite générique d'arrivée de ces messages doit être très importante en cours de collecte. En effet, un producteur peut bloquer l'accès aux robots de la BnF si le DLN ne réagit pas rapidement[100].

Toutes les modifications et vérifications effectuées par un agent du DLN sont notées dans un journal de bord afin de tenir au courant les collègues et de garder une trace des actions réalisées.

3. Contrôle de la qualité des collectes

La surveillance des collectes en cours est déjà en soit une étape de contrôle qualité. En effet, la possibilité d'ajouter du budget ou encore de corriger certains URL permet d'assurer une meilleure collecte.

[100] Il est possible de modifier les paramètres de collecte pour éviter la surcharge du site collecté par exemple.

Un outil de contrôle de réception élaboré par le DSI produits des indicateurs à la fin de chaque capture (nombre d'URL collecté, poids de la collecte, nombre de code 200…).

Le contrôle de la qualité des collectes, effectué une fois que les données ont été indexées et sont disponibles dans l'interface d'accès, se fait sous la responsabilité des correspondants. La vérification visuelle de la qualité des URL récoltées se fait à l'aide de l'interface d'accès Archives de l'Internet[101]. Il est également possible d'effectuer des vérifications plus précises avec l'application WayBack Machine[102], en utilisant la recherche avancée.

Le contrôle visuel ne peut s'effectuer que sur un échantillon de sites au vu de la masse qui est collectée. Cet échantillon est constitué :

- des sites non collectés et redirigés, repérés à partir du « seed-report » de la fiche job NAS ;
- de sites d'extensions variées, de typologies différentes (sites institutionnels, blogs…), de sites avec des architectures variées ;
- de sites de grande taille, repérés à partir de l'outil de contrôle de réception ;
- des cas particuliers relevés lors de la collecte (indiqué dans le journal de bord) ;
- des sites qui ont posé problème lors de la collecte ou lors des collectes précédentes.

Ce contrôle visuel doit être effectué avant le départ de la collecte suivante afin d'y intégrer les modifications nécessaires.

Le contrôle visuel consiste à vérifier :

- la maquette du site en la comparant à celle présente sur le Web vivant ;
- la profondeur de collecte en accédant aux pages les plus profondes par clics successifs ;
- la présence du contenu visé par le sélectionneur.

Les résultats du contrôle sont reportés dans BCWeb.

[101] Voir en annexe les captures d'écrans relatifs à l'interface d'accès.

[102] Il s'agit de l'interface d'accès pour les professionnels.

4. Etude préalable avant le lancement d'une nouvelle collecte ciblée

Durant le stage, la mission d'élaborer une nouvelle collecte ciblée m'a également été confiée. Cette collecte ciblée concernait les plateformes de blogs. En fait, les blogs sont assez mal collectés par les collectes larges[103]. Celles-ci ne sont pas assez profondes pour permettre de récupérer un nombre signifiant de blogs. Il fallait donc trouver une méthode qui permettrait de capturer un nombre signifiant de blogs, sans préjuger de la qualité ou de l'intérêt documentaire des contenus. En effet, en tant qu'établissement dépositaire du dépôt légal du Web, la BnF doit pouvoir collecter un maximum de blogs français, représentatifs de l'ensemble de la blogosphère française, dans la mesure du possible.

D'emblée, la possibilité de sélectionner un à un les blogs et les entrer dans BCWeb en créant une collecte « blogs », a été écartée. En effet, il aurait été impossible d'obtenir un nombre important de blogs par cette méthode.

Tout d'abord, nous avons identifié les plateformes qui ont fourni le plus grand nombre d'URL lors de la collecte large de 2011. Il fallait ensuite établir une liste d'hôtes[104] la plus large possible, afin qu'elle soit la plus représentative des blogs français. Cette liste a été constituée à partir de plusieurs sources :

- Les ingénieurs de production ont extrait les différents hôtes collectés des plateformes choisies, à partir des données produites par l'ensemble des collectes du Web menées par la BnF depuis 2010.
- Une liste d'hôtes a été établie en lançant une requête sur le moteur de recherche Google. Les cent premiers résultats ont été extraits grâce à un outil disponible en ligne[105]. Cette solution a l'avantage de ne cibler que les blogs les plus consultés, ce qui peut représenter un bon échantillon du paysage des blogs français.
- Une troisième liste a été établie grâce aux annuaires proposés par les plateformes elles-mêmes. Ces derniers répertorient les cent ou deux cents blogs les plus consultés.

[103] BCWeb contient également des blogs en collecte ciblée, ces derniers ont été sélectionnés en fonction de leur pertinence, par rapport à une collecte donnée (exemple du blog d'un militant dans la collecte Elections).

[104] Un hôte hébergé sur une plateforme correspond à un blog.

[105] Le site http://outils-seo.alwaysdata.net propose plusieurs outils dédiés au référencement des pages web. L'un de ces outils permet d'extraire les url des résultats de google.fr suivant une requête (http://outils-seo.alwaysdata.net/outils-traitement-analyse/extraire-url-resultat-google/), consulté le 14 septembre 2012.

Les trois listes ont été fusionnées et dédoublonnées. Des tests de collecte ont également été menés au mois de juillet pour identifier les blogs qui n'étaient plus en ligne. Ces tests ont permis de constituer une liste finale d'URL entièrement valide[106]. Par ailleurs, il fallait déterminer des paramètres de collecte spécifiques, en collaboration avec les ingénieurs de production. En effet, étant donné que plusieurs hôtes allaient être collectés sur un même domaine, une attention particulière devait être portée à ne pas surcharger les serveurs des plateformes collectées.

C. Valorisation du projet et des collections constituées auprès de chercheurs et de partenaires

La collecte du Web électoral constitue désormais une collection de sources primaires sur une période de dix ans. Il est indispensable d'effectuer des opérations de valorisation afin de faire connaître cette collection. Cependant, s'agissant d'une collecte acquise par la voie du dépôt légal, il n'est pas possible de mettre à la disposition du public les sites archivés, en dehors des espaces recherche de la Bibliothèque. La valorisation de la collection vers un public plus large que celui des chercheurs, se retrouve donc bridée par cette condition.

Concernant la valorisation de la collection du Web électoral vers le public des chercheurs, la BnF peut promouvoir la connaissance et l'usage de cette collection en enrichissant par exemple, le parcours guidé consacré aux élections. La Bibliothèque pourrait également prévoir la création d'un autre parcours guidé illustrant par exemple uniquement les élections présidentielles depuis 2002. Des approches plus locales peuvent également être envisagées, dans ce cas l'idée d'un partenariat avec les BDLI autour de la valorisation prendrait tout son sens. De plus, cette forme de valorisation sera plus facile à mettre en place une fois que les BDLI auront un accès à distance aux collections des archives du Web. Mettre en place cet accès à distance dans les BDLI (accès toujours restreint aux lecteurs disposant d'une accréditation) est en soi une opération de valorisation.

D'autre part, deux projets de valorisation ont vu le jour durant la phase de collecte du Web électoral 2012. La première consiste à mettre en ligne l'intégralité des adresses de sites archivés dans le cadre des collectes du Web électoral depuis 2002. La seconde consiste à mettre en ligne sur le site de la BnF, un bilan du projet Elections 2012.

[106] Au final, une liste de près de 450 000 URL de départ valides a été obtenue, ce qui est considérable pour une collecte ciblée.

1. Participation à la mise en ligne des données sur data.gouv.fr

Le service du dépôt légal numérique a mis en ligne sur la plateforme data.gouv.fr[107], la liste des sites électoraux archivés depuis 2002[108].

Les données sont sous forme de fichiers CSV regroupant toutes les URL collectées entre 2002 et 2010, sachant que les données de 2012 les rejoindront par la suite. Ces URL sont associées à diverses informations d'ordre documentaire (région, type de site, parti) et technique (fréquence et profondeur de collecte) lorsque celles-ci sont disponibles. Il a donc fallu extraire les données, les harmoniser et les organiser de manière à les rendre utilisables par les chercheurs mais également par un public plus large. En effet, l'objectif de cette opération est double. D'une part, cette opération permet de valoriser la collection du Web électoral et d'inciter les chercheurs à venir les consulter en salles de lecture. D'autre part, elle permet de fournir aux journalistes et aux chercheurs des données permettant de travailler sur le Web électoral : évaluer le nombre de sites, la répartition des sites par parti politique ou encore calculer le taux de disparition. Cette publication participe également à la démarche d'ouverture des données publiques à la BnF.

2. Participation à la rédaction d'un bilan sur le projet Elections 2012

Depuis 2002, chaque collecte du Web électoral a été documentée par un document interne à la BnF. Ce document dresse le bilan de la collecte aussi bien sur le plan documentaire (analyse du contenu des sites, leur utilité…) que sur le plan organisationnel (organisation pratique du travail mené par les agents de la BnF et des BDLI). D'un point de vue documentaire, ce bilan remplit plusieurs objectifs :

- Il permet de conserver la mémoire des choix documentaires effectués, dans la perspective de futurs projets événementiels.
- Il permet également de caractériser la collection (connaître la nature des fichiers collectés : images, vidéos, texte…)

[107] Data.Gouv.fr est une plateforme de diffusion de données publiques ouvertes à la réutilisation dans les termes et les conditions de la Licence Ouverte (Open Licence). Les contenus éditoriaux de data.gouv.fr sont placés sous Licence Ouverte.

[108] La liste peut être téléchargée à cette adresse : http://www.data.gouv.fr/donnees/view/Collectes-du-Web-électoral-par-la-BnF-551866 (consulté le 14 novembre 2012)

- Ce document permet de décrire le projet, en direction des futurs chercheurs. Il s'agit d'expliquer comment s'est constituée cette collection, dans une perspective de documentation du contexte de production. Il est indispensable d'expliquer les limites qui ont pu être rencontrées et les choix qui ont été effectués.
- Enfin, le document apporte également un éclairage sur la Web campagne en proposant des pistes de recherche pour explorer l'ensemble documentaire constitué par la BnF.

Sur le plan organisationnel, ce bilan permet d'évaluer le coût du projet (financier, matériel et humain), de redéfinir en conséquence l'organisation des futurs collectes dans le cadre du « modèle intégré » de la BnF. Enfin, ce bilan permet de partager l'expérience de la collecte au niveau national (vers les BDLI) et international (IIPC). Comme nous l'avons vu précédemment, le partage d'expérience est l'un des objectifs à réaliser au sein du Consortium IIPC.

Cette année, il est envisagé de mettre en ligne ce bilan de la collecte sur le site de la BnF afin que le plus grand nombre y ait accès.

CONCLUSION

Ce stage de six mois a été une véritable première expérience professionnelle. Cette expérience s'est avérée très positive sur différents plans.

D'abord, le plan d'accueil qui m'a été proposé (programme de visites et d'échanges au sein de la Direction des Collections et de la Direction des Services et des réseaux) m'a permis de découvrir la diversité des activités exercées par l'établissement. Ces visites m'ont également permis de voir jusqu'à quel point le dépôt légal du Web a été intégré au sein de la BnF. En effet, l'archivage du Web n'est pas le travail d'un seul service, c'est une activité qui implique beaucoup de personnes, chacune ayant son propre domaine de compétence. La diversité des témoignages recueillis m'a permis de confronter leurs expériences, et de voir à quel point il était nécessaire d'avoir des intermédiaires entre les bibliothécaires et les ingénieurs. En fait, il était très intéressant de voir comment la logique des bibliothécaires divergeait souvent de celle des ingénieurs. Il était donc important que le service du Dépôt légal numérique joue ce rôle d'intermédiaire entre les deux. Afin de comprendre les problèmes de chaque parti et jouer ce rôle d'interface, les agents du DLN doivent faire preuve de compétences hybrides. Ce stage m'a donc permis de développer des connaissances documentaires mais aussi techniques. Travailler dans un service qui associe des compétences d'archivistiques à des compétences informatiques a donc été une expérience très enrichissante. Il fallait constamment trouver des solutions techniques tout en faisant attention aux problématiques patrimoniales.

Sur le plan humain, le travail au sein d'un petit service m'a paru intéressant et agréable. J'ai en effet apprécié le cadre de travail et les moyens mis à disposition du personnel, ainsi que les relations humaines enrichissantes.

D'autre part, participer à une activité de portée internationale m'a permis de comprendre comment pouvait fonctionner un système de collaboration à l'échelle de la planète.

BIBLIOGRAPHIE

Ouvrage :

Cardon, Dominique, *La démocratie internet. Promesses et limites*, Paris, Seuil, 2010

Chevallier, Philippe, Illien, Gildas, *Les Archives de l'internet. Une étude prospective sur les représentations et les attentes des utilisateurs potentiels*, Paris, BnF, 2011, consultable en ligne : http://www.bnf.fr/documents/enquete_archives_web.pdf, consulté le 10 septembre 2012

Ouvrage collectif :

Greffet, Fabienne (dir.), *Continuerlalutte.com. Les partis politiques sur le web*, Paris, Les presses Sciences Po, 2011

Contribution dans un ouvrage collectif :

Illien, Gildas, Oury, Clément, « Quelle politique documentaire pour l'archivage des sites Internet ? » dans Carbone, Pierre et Cavalier, François (dir.), *Les collections électroniques, une politique documentaire en mouvement*, Paris : Éditions du Cercle de la librairie, 2009, p. 157-178

Article dans une revue :

Brügger, Niels, « L'historiographie de sites Web : quelques enjeux fondamentaux », *Le Temps des médias*, 2012/1 n° 18, 2012, p. 159-169

Greffet, Fabienne, « Le web dans la recherche en science politique. Nouveaux terrains, nouveaux enjeux », *Revue de la Bibliothèque nationale de France*, Paris, n° 40, 2012, p. 78-83

Illien, Gildas, « Le dépôt légal de l'Internet en pratique », *BBF*, 2008, n° 6, p. 20-27 Consultable en ligne : http://bbf.enssib.fr/, consulté le 4 septembre 2012

Illien, Gildas, « Une histoire politique de l'archivage du web », *BBF*, 2011, n° 2, p. 60-68 Consultable en ligne : http://bbf.enssib.fr/, consulté le 07 septembre 2012

Oury, Clément, « Soixante millions de fichiers pour un scrutin, les collections de sites politiques à la BnF », *Revue de la Bibliothèque nationale de France*, Paris, n° 40, 2012, p. 84-90

Article de presse :

Berne, Xavier, « Dans les coulisses du dépôt légal de l'Internet », *PC INpact*, Paris, 14 juin 2012, consultable en ligne : http://www.pcinpact.com/dossier/depot-legal-internet-bnf archives/1.htm, consulté le 28 août 2012

Intervention :

Bermès, Emmanuelle, Dussert Carbon, Isabelle, Ledoux, Thomas et Ludovici, Christian, « La préservation numérique à la Bibliothèque nationale de France : présentation technique et organisationnelle », intervention lors du congrès de l'IFLA en 2008, disponible à l'adresse : http://archive.ifla.org/IV/ifla74/papers/084-Bermes_Carbone_Ledoux_Lupovici-trans-fr.pdf, consulté le 7 septembre 2012

Textes de loi :

Loi n° 2006-961 du 1er août 2006 sur le droit d'auteur et les droits voisins dans la société de l'information (DADVSI). Disponible à l'adresse :
http://www.legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT000000266350

Code du patrimoine, titre III sur le dépôt légal, art. L. art. 131-1 à 133-1. Disponible à l'adresse : http://www.legifrance.gouv.fr/affichCode.do?cidTexte=LEGITEXT000006074236

Décret n° 2006-696 du 13 juin 2006 modifiant le décret n° 93-1429 du 31 décembre 1993 relatif au dépôt légal. Disponible à l'adresse :
http://www.legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT000000816574&dateTexte=

Site Internet :

http://www.bnf.fr/fr/professionnels/depot_legal.html
Site de la BnF, pages consacrées au dépôt légal
Informations pratiques et textes de références sur le dépôt légal,
Consulté le 14 septembre 2012.

http:// www.bnf.fr/fr/professionnels/conserver_spar.html
Site de la BnF, pages consacrés au projet SPAR
Présentation, infrastructure, réalisation de la partie logicielle
Consulté le 7 septembre 2012

http://www.data.gouv.fr/donnees/view/Collectes-du-Web-électoral-par-la-BnF-551866
Plateforme de diffusion de données publiques
Listes des URL collectées lors des collectes électorales menées par la BnF
Consulté le 15 septembre 2012

http://www.Webarchive.org.uk/ukwa/
Site de la *British Library*
Accès aux collections des archives du Web, informations générales
Consulté le 7 septembre 2012

http://www.netpreserve.org/
Site d'IIPC (International Internet Preservation Consortium)
Informations sur l'organisme et sur ses membres
Consulté le 7 septembre 2012

http://archive.org/
Site d'Internet Archive
Accès aux collections d'archives du Web, informations pratiques sur les logiciels *open source*
Consulté le 14 septembre 2012

Document interne à la BnF :

« Elections 2007 : Bilan documentaire du projet »

«Le dépôt légal de l'Internet électoral en 2012. Journée de travail et de formation avec les BDLI partenaires »

« Memento de production : formalisation des opérations internes au DLN, année 2012 »

TABLES DES ANNEXES

Annexe 1 :

Organigramme de la Bibliothèque nationale de France

mise à jour 2 février 2012

Président
Bruno RACINE

chargée de mission
Christine SEMENCE

Conseil d'administration
président : Bruno RACINE

Conseil scientifique
président : Roger CHARTIER

Contrôleur financier
Florence SEVIN-DAVIES

Agence comptable
agent comptable : Etienne EFFA
fondée de pouvoir : Carmen BEDOU
service de Paiement des fournisseurs
Maryvonne BURATTI
service des Dépenses de personnel
Fabienne MOREAU
service Comptabilité recette
Corinne GORGIN

directrice de projet filiale BnF
Nathalie THOUNY

chef de la Mission de communication interne
Catherine PETIT

chargée de la Mission handicap
Carole ROUX

chef de la Mission pour la gestion de la production documentaire et des archives
Aurélien CONRAUX

Unité des Sapeurs-Pompiers de Paris
Capitaine Sébastien ALBERTINI

inspecteur Hygiène et sécurité
Oreste FRISCIRA

Directrice générale
Jacqueline SANSON

directeurs généraux adjoints
Arnaud BEAUFORT
Denis BRUCKMANN
Sophie DURRLEMAN

Délégation à la Diffusion culturelle

délégué : Thierry GRILLET
adjointe : Cécile PORTIER
chargé de mission : Bruno BAUDRY

service des Expositions : Ariane JAMES-SARAZIN
service de Coordination générale
Stéphane DUCHESNE
service des Manifestations : Jean-Loup GRATON
service de l'Action pédagogique
Lucile TRUNEL
service de l'Accueil général : Pierre CIER
Département Éditorial et commercial
Jocelyn RIGAULT
service Commercial : Christophe STOOP
service de l'Édition des livres
Pierrette CROUZET-DAURAT
service des Éditions multimédias
Françoise JUHEL

Délégation aux Relations internationales

délégué : Gérald GRUNBERG
adjoint et chargé de mission
Europe/coopération bilatérale
Bruno SAGNA
chargés de mission
CENL
Karin VARLOOT-KNUDSEN
Europe/coopération multilatérale
Elisabeth FREYRE
Francophonie
Franck HURINVILLE
Amérique Asie
Mireille BALLIT
directeur du programme IFLA/PAC
Christiane BARYLA

Délégation à la Communication

délégué : Marc RASSAT
adjointes : Françoise GUILLERMO,
Claudine HERMABESSIERE
service de Presse : Claudine HERMABESSIERE
Éditorial : Sylvie LISIECKI
service de Coordination Internet
Hélène TOUTUT
Graphisme : Françoise TANNIÈRES

Délégation à la Stratégie et à la recherche

délégué : Thierry PARDÉ
chargés de mission
Julien BARBIER
Philippe CHEVALLIER
Olivier JACQUOT
Véronique MICHEL
Cécile TOUITOU

Délégation au Mécénat

déléguée : Kara LENNON CASANOVA

Direction des Collections

directeur : Denis BRUCKMANN

adjoint chargé des Affaires administratives et financières : Georges-Henri VERGNE
service des Affaires budgétaires : Nathalie COHIN
service de Coordination des ressources : N.
service de la Gestion centralisée des acquisitions
Françoise CHAUVIÈRE

adjointe chargée des Affaires scientifiques et techniques, chef de la Mission de coordination :
Isabelle LE MASNE DE CHERMONT
service de l'Orientation des lecteurs
Marie-Hélène PONS
service de l'Inventaire rétrospectif
Jean-Dominique MELLOT

chef du projet Richelieu : Marie DE LAUBIER

chef de projet Évolution du Haut-de-jardin
Isabelle MANGOU

Bibliothèque de l'Arsenal
directeur : Bruno BLASSELLE
adjointe : Ève NETCHINE
Collections : Claire LESAGE
Communication et conservation : Sophie NAWROCKI

Département des Arts du spectacle
directeur : Joël HUTHWOHL
adjointe : Véronique MEUNIER
Archives, imprimés : Joëlle GARCIA
Iconographie, documentation : Caroline RAYNAUD
Conservation, communication
Marie-Odile ILLIANO
Maison Jean-Vilar (Avignon)
Lenka BOKOVA

Département des Cartes et plans
directeur : Jean-Yves SARAZIN
adjoint : François NAWROCKI
Dépôt légal et fonds patrimoniaux
Catherine HOFMANN
Acquisitions et collections géographiques
Olivier LOISEAUX
Conservation et communication
Benoît COTE COLISSON

Département des Estampes et de la photographie
directrice : Sylvie AUBENAS
adjointe : Corinne LE BITOUZÉ
Estampe ancienne et de la réserve
Séverine LEPAPE
Estampe moderne et contemporaine, affiche, image
Anne-Marie SAUVAGE
Photographie : Anne BIROLEAU-LEMAGNY
Conservation et communication : Loïc LE BAIL

Département des Manuscrits
directrice
Isabelle LE MASNE DE CHERMONT (intérim)
adjointe : Anne-Sophie DELHAYE
Manuscrits médiévaux : Charlotte DENOËL
Manuscrits modernes et contemporains
Marie-Laure PREVOST
Manuscrits orientaux : Laurent HERICHER
Conservation et communication
Pierre-Jean RIAMOND

Département des Monnaies, médailles et antiques
directeur : Michel AMANDRY
adjointe : Marielle PIC

Département de la Musique
directrice : Elizabeth GIULIANI
adjoint : Michel YVON
Collections patrimoniales : Laurence DECOBERT
Collections imprimées : Marguerite SABLONNIERE
Conservation et communication
Catherine VALLET-COLLOT
Bibliothèque-musée de l'Opéra
directeur : Pierre VIDAL

Département de la Recherche bibliographique
directeur : Raymond-Josué SECKEL
adjointe : Catherine ÉLOI

Département Philosophie, histoire, sciences de l'homme
directrice : Claude COLLARD
adjointe : N.
Philosophie/Religion : Valerie GRIGNOUX
Histoire : Frédéric MANFRIN
Sciences sociales : Cristina ION
Conservation : Bruno JEANNET

Département Droit, économie, politique
directeur : Pascal SANZ
adjointe : Lise DEVREUX
Droit, Science politique : Gilles BAUDOUIN
Economie, PRISME : Sébastien BOGAERT
Presse : Philippe MEZZASALMA
Conservation : Cécile FORMAGLIO

Département Sciences et techniques
directeur : Hervé COLINMAIRE
adjointe : Florence USCLAT
Histoire des sciences, sciences fondamentales
Angel CLEMARES
Sciences de la vie, sciences de l'ingénieur
Eva FUENTES
Conservation : Martine DELAVEAU

Département Littérature et art
directeur : Jean-Marie COMPTE
adjointe : Anne PASQUIGNON
Centre national de littérature pour la jeunesse – La Joie par les livres
directeur : Jacques VIDAL-NAQUET
Littérature française : Laurence RATIER
Littératures étrangères : Catherine AURERIN
Littératures orientales et art : Cheng PEI
Documentation sur le livre et la lecture
Bérénice STOLL
Conservation et entrées : Olivier SIFFRIN

Département de l'Audiovisuel
directrice : Isabelle GIANNATTASIO
adjoint : N.
service Public : Régine LARTIGUE
Documents sonores : Pascal CORDEREIX
Images : Alain CAROU
Multimédia : Xavier SENE
Consultation et conservation : Pierre ECHEGARAY

Réserve des Livres rares
directeur : Antoine CORON
adjointe : Geneviève GUILLEMINOT

Direction des Services et des réseaux

directeur : Arnaud BEAUFORT
adjointe chargée des questions scientifiques et techniques : Catherine DHÉRENT

Mission de coordination administrative et financière : Mireille NOUVEL

Département Information bibliographique et numérique
directeur : Gildas ILLIEN
adjointe : Françoise BOURDON
service Coordination et administration des métadonnées : Jérôme VILLEMINOZ
service Prospective et services documentaires
Marianne CLATIN

Département de la Conservation
directrice : Isabelle DUSSERT-CARBONE
adjoint : Philippe VALLAS
service Central : Philippe VALLAS
service de Numérisation : Laurent DUPLOUY
service Technique du site François-Mitterrand
Patrice RACT
Centre technique de Bussy-Saint-Georges
Gilles BEDDOK
service de Restauration
Isabelle ROLLET
Centre de conservation Joël-Le-Theule (Sablé-sur-Sarthe)
Annie BONNAUD

Département de la Coopération
directrice : Aline GIRARD
service Catalogue collectif de France [CCFr]
Véronique FALCONNET
service Pôles associés – Gallica
Frédéric MARTIN

Département du Dépôt légal
directrice : Hélène JACOBSEN
adjoint et service Redistribution documentaire
Benoît TULEU
Bibliographie nationale française/périodiques
Dominique BRÉCHIROUX
Bibliographie nationale française/livres
Jean-Charles PAJOU
service du Dépôt légal numérique
Clément OURY
service de la Gestion des périodiques
Alix BRUYS
service de la Gestion des livres : Agnès DEMAY

Département de la Reproduction
directeur : Sébastien PETRATOS
adjoint : Yves LE GUILLOU
service Clients : Hélène GUY LHOMME
service Production : Patrick BRAMOULLÉ
service Redevance comptabilité : Éric MAGNIEN

Département des Systèmes d'information
directeur : Jean-Marc CZAPLINSKI
adjoint : Adotè CHILLOH
service Support et production : Adotè CHILLOH
service Études et développement
Catherine NEUFOND
Mission coordination et projets : Serge PIZZIO

Direction de l'Administration et du personnel

directrice : Sophie DURRLEMAN
adjoint : Pierre-Henry COLOMBIER

Chargée de la valorisation économique
Alix LAME-BERGIS

Direction déléguée aux Ressources humaines
directeur : Bertrand WALLON

service de la Formation et des qualifications
Michel NETZER
service Logistique des ressources humaines
Marianne PICART
service de l'Action sociale : Bénédicte JACOB
service de Médecine de prévention
Docteur Isabelle ARNAUD-VERGNIOL
Psychologue : Monique DUJARDIN

Département du Personnel et de l'emploi
directrice : Noëlle BARILLAT-DERLON
service de l'Emploi et des crédits
Anne-Sophie DE BELLEGARDE
service de l'Administration du personnel
Danièle SAUTROT

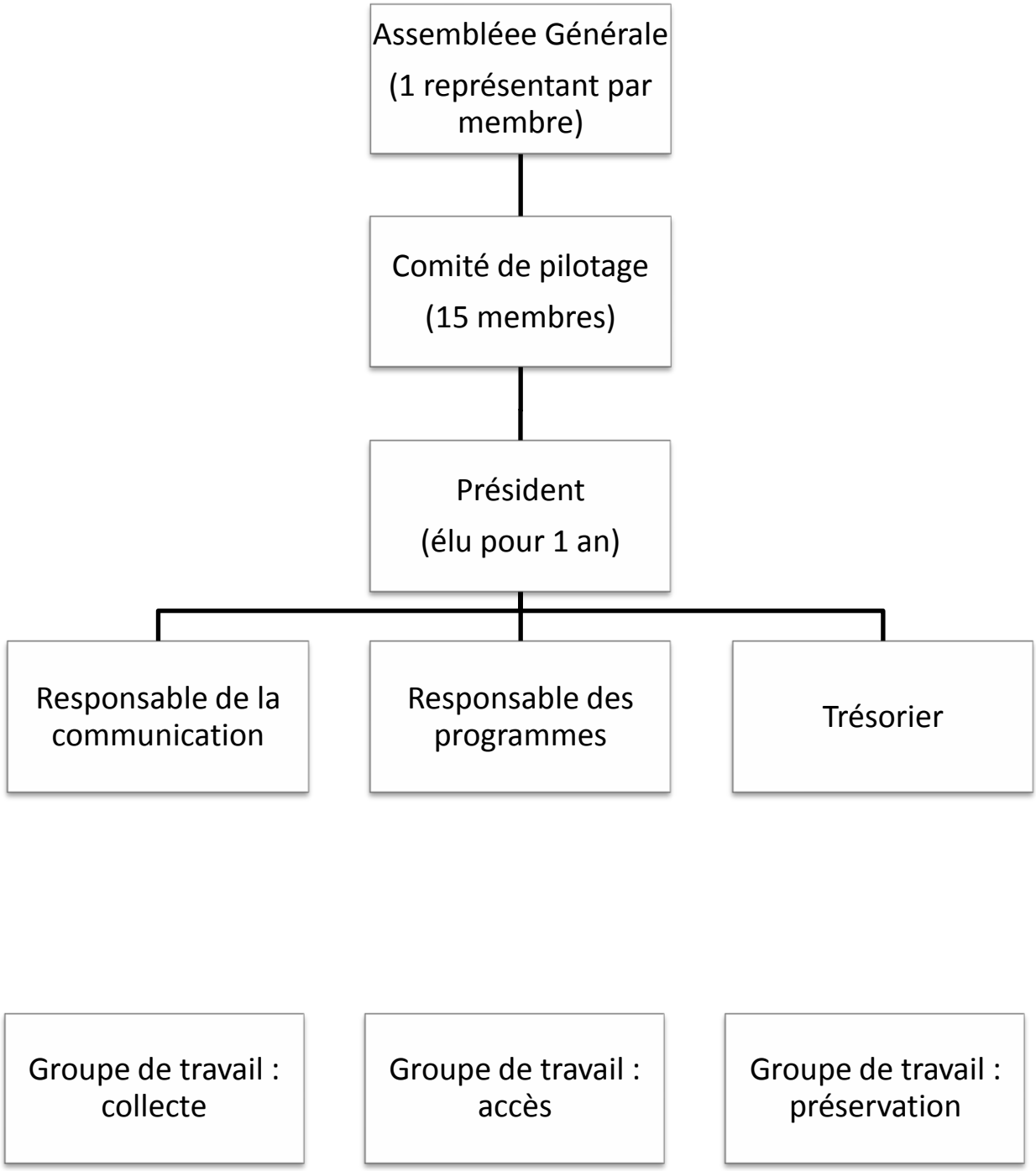
Département des Affaires juridiques
directeur : Stéphane ALCANDRE
adjointe : Sophie SEPETJAN
service des Marchés : Stéphane ALCANDRE
service Juridique : Sophie SEPETJAN

Département du Budget et des affaires financières
directrice : Mireille FATON
adjoint : Jamal BOUTOUMI
service de l'Informatique financière
Jamal BOUTOUMI
service du Budget
Corinne GRANGE
service de l'Ordonnancement et des crédits déconcentrés
Isabelle EDET
service Financier de la direction de l'administration et du personnel
Françoise PEREIRA
service de la Billetterie
Alain JUMEL

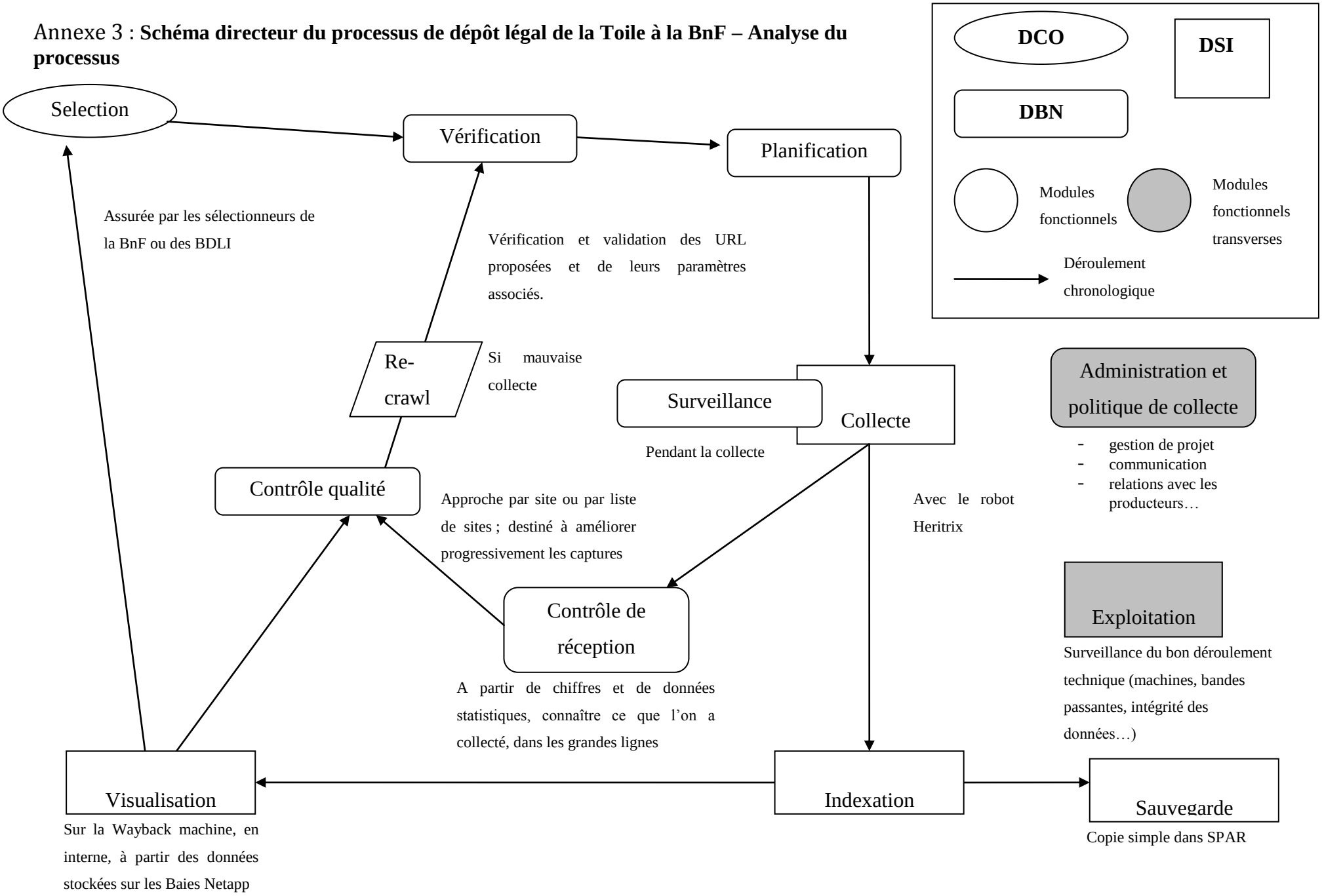
Département des Moyens techniques
directeur : Pierre-Henry COLOMBIER
adjoint : Daniel DURRITCAGUE
Division Travaux Maintenance Exploitation
Philippe ROUSSEAU
– service Lots architecturaux
Adalberto TERCEROS
– service Courants faibles : Daniel JAOUEN
– service Courants forts
François HORT-VISSETH (intérim)
– service Climatisation : Olivier BARMA
– Intendance Richelieu
Cécile DESCAMPS FILIATRE
– Intendance Bussy-Saint-Georges
François PERNOIS
service Hygiène - sécurité - environnement
Jean-Luc DURAND
service Logistique et moyens généraux
Pierre BOYER
service de la Sûreté : Jérome RESIBEAU

Annexe 2 :

Organigramme : le fonctionnement d'IIPC



Annexe 3 : **Schéma directeur du processus de dépôt légal de la Toile à la BnF – Analyse du processus**



Les tâches assurées par l'équipe du DL Web :

Sélection : Il s'agit de la tâche assurée par les sélectionneurs.

Vérification et validation : les propositions de sites à capturer font l'objet d'une vérification technique par les agents du DLN. Il s'agit de vérifier que les sites de la liste sont techniquement capturables (des animations flash, des formulaires de requête, la présence de # dans les url peuvent en effet constituer un obstacle pour le robot), que les URL de départ spécifiées sont les plus pertinentes et que les paramètres de fréquence et de profondeur choisis sont cohérents par rapport au contenu du site.

Planification : La planification s'articule autour deux tâches distinctes : la création et la gestion du calendrier de collecte, et la préparation des collectes.

Les collectes s'effectuent à partir d'un calendrier de collecte établi au sein du DLN. Ce calendrier est constitué à partir des fréquences de collecte des sites à capturer. La gestion du planning veille à la bonne organisation des collectes afin qu'elles puissent être réalisées selon le calendrier pré-établi. Plusieurs critères interviennent : le nombre de sites à collecter, les temps de capture estimés de chaque site, le moment de la collecte (début, fin de semaine) et la disponibilité des serveurs de collecte.

La préparation des collectes consiste à basculer la liste des sites à collecter de l'outil de proposition (BCWeb) vers l'outil de planification des collectes (NAS).

Collecte : la liste des URL graines à collecter ce jour est répartie en un certain nombre de *jobs*, en fonction de la profondeur de collecte et de la taille des sites. Chaque ordinateur de collecte, sur lequel est installée une instance du robot Heritrix, se voit attribuer un certain nombre de *jobs* à traiter.

Chaque robot est alors lancé sur la liste des URL graines du *job* en cours. Le robot fonctionne en quelque sorte comme un internaute automatique. Il visite tout d'abord la page à laquelle correspond une URL de départ. Il identifie, au sein de cette page, tous les liens hypertexte pointant vers d'autres fichiers (au sein de la même page Web, au sein du même site, ou dans un autre site). Il inscrit, s'il convient, l'adresse de ces fichiers dans une liste d'attente (la « queue »). Enfin, il archive la page qu'il vient de visiter et de traiter, en la copiant au sein d'un fichier ARC.

Après avoir collecté toutes les URL graines, le robot capture les fichiers correspondant aux URL placées dans la liste d'attente, qui à leur tour vont livrer de nouvelles URL à collecter… Ceci va durer jusqu'à ce que le robot ait découvert et archivé toutes les URL pertinentes.

De fait, tous les fichiers découverts par le robot ne sont pas placés dans la liste d'attente. Lorsqu'un fichier est découvert, le robot vérifie s'il appartient au périmètre (*scope*) retenu. Ce périmètre peut être restreint, dans le cadre de collectes ciblées (le robot peut par exemple ne collecter que les fichiers appartenant à un site donné, c'est-à-dire ne suivre que les liens internes au sein de ce site : c'était le cas

pour la campagne d'archivage des sites électoraux). Le périmètre peut être beaucoup plus étendu, dans le cadre de collectes larges à vocation exploratoire (par exemple, accepter toutes les URL en « .fr »).

Surveillance : Lors de la collecte, il est nécessaire de surveiller très attentivement l'activité du robot. Le robot peut facilement s'enferrer dans des pièges, c'est-à-dire des pages présentant des liens vers des URL vides ou redondantes, que le robot peut archiver à l'infini.

Les calendriers présents sur les blogs présentent un exemple caractéristique. La plupart du temps, les jours affichés par ces calendriers sont cliquables, afin que l'internaute puisse accéder au message posté le jour en question. Toutefois, certains calendriers permettent de cliquer sur tous les jours, y compris ceux pour lesquels aucun message n'a été posté ; ces calendriers permettent également de s'étendre à l'infini dans le temps (on peut cliquer jusqu'à l'an 3000…). A chaque clic, une URL différente est générée, mais le message affiché sur la page est toujours du type : « Aucun message pour ce jour ».

Si cet exemple est simple à se représenter et assez significatif, il n'épuise pas la multitude des pièges que peut rencontrer le robot : certains sites donnent un numéro de session à chaque connexion, numéro de session qui se retrouve dans l'URL, qui fait donc croire au robot qu'il a affaire à une URL différente à chaque connexion. Parfois, c'est le robot lui-même, en archivant des javascripts, qui crée de fausses URL…

Pour des raisons évidentes de stockage, de gestion du temps et du coût de production, toutes ces informations inutiles ne doivent pas être collectées. La surveillance attentive des journaux d'événements d'Heritrix permet d'identifier les pièges et de les traiter. Les mauvaises URL sont extraites de la collecte, des filtres spécifiques sont créés afin d'aider Heritrix à identifier et à éliminer automatiquement les mauvaises URL. Certains paramètres techniques sont ajustés (nombre maximum de liens que le robot est autorisé à suivre…).

Contrôle de réception : L'objectif du contrôle de réception est de vérifier si la commande a bien été réalisée et d'en dresser un premier bilan chiffré. La vérification de la commande consiste à s'assurer, à partir des différents rapports d'Heritrix, que toutes les URL contenues dans la liste de départ (les *seeds*) ont bien été collectées, et d'isoler les URL non collectées pour analyse. Le bilan chiffré permet ainsi de fournir à l'état brut des indicateurs quantitatifs afin de donner des informations claires sur la collecte réalisée. Les indicateurs comme le nombre d'URL capturées, la durée de la collecte, la répartition par types MIME sont autant d'indices du bon ou du mauvais déroulement de la collecte. Ces indicateurs sont tous attachés à un *job* particulier. Lorsque ce *job* est lancé de façon récurrente, ils permettent de suivre l'évolution de la taille et de la forme des sites contenus dans le *job*.

Ces indicateurs servent aussi à la planification dans la gestion et l'organisation du calendrier de collecte.

Contrôle qualité : Le contrôle qualité intervient à plusieurs moments dans la chaîne de production.

Un premier contrôle est effectué après collecte au moyen des rapports et des journaux d'événements produits par le robot de collecte. A l'aide des indicateurs, il permet de vérifier pour chaque *job* ou pour chaque site que la collecte a été correctement réalisée. Aussi par exemple, si le nombre d'URL collectées diminue ou s'effondre, c'est le signe qu'un problème est survenu lors de la collecte. Cela a été le cas des pages Facebook qui, au cours du projet, ont refondus complètement l'architecture technique de leur site. Le robot ne parvenait pas à capturer la nouvelle version, problème qui a été réglé grâce à une instruction technique.

Un deuxième contrôle, le contrôle visuel, est effectué après indexation, site par site. Il permet de vérifier la qualité réelle de la collecte, sa lisibilité, sa navigabilité, sa profondeur. Ce contrôle site par site est effectué par échantillonnage. Lorsqu'un site est collecté pour la première fois, il fait systématiquement l'objet d'un contrôle visuel.

Les problèmes de collecte identifiés lors de ces contrôles sont signifiés à la validation technique, qui pourra changer les paramètres de collecte. Ils sont aussi notifiés à l'ingénieur de collecte, s'il est besoin d'y trouver des solutions techniques.

Indexation : Après leur collecte, les fichiers (stockés au sein de fichiers ARC) sont envoyés sur un serveur d'indexation. L'indexation consiste, à partir des fichiers ARC, à produire trois types de fichiers :

- Les fichiers DAT sont destinés à retrouver les enregistrements ARC au sein d'un fichier ARC. Ils sont constitués d'un en-tête DAT, suivis d'une série de blocs de données, chacun correspondant à un enregistrement ARC. Chaque bloc de données reprend certaines informations contenues dans l'enregistrement ARC auquel il correspond, notamment son URL, son empreinte numérique, puis son emplacement au sein du fichier ARC (c'est-à-dire l'octet à partir duquel l'enregistrement ARC commence, et la taille de l'enregistrement en octets). Enfin, si le bloc correspond à l'enregistrement d'un fichier HTML, il contient la structure des pages HTML, c'est-à-dire une liste des liens sortants.
- Les fichiers CDX sont construits à partir des fichiers DAT. Il s'agit d'index triés ligne par ligne vers des enregistrements ARC. Lorsqu'une URL fait l'objet d'une requête (de la part d'une interface d'accès, par exemple), l'index CDX permet d'indiquer dans quel fichier ARC et à quelle place dans ce fichier ARC se trouve le fichier auquel correspond l'URL.
- Enfin, les fichiers Path servent à retrouver sur quel support physique a été stocké un fichier ARC.

L'objectif de l'indexation est de rendre accessible les données. L'ensemble des données collectées dans le cadre des élections sont rassemblées sous un seul index (nommé elections2012).

Visualisation : Tous les fichiers produits sont transférés sur des serveurs d'accès disposant d'une très grande capacité de stockage, les baies Netapp. Ce transfert se fait en continu, au fur et à mesure que le travail d'indexation s'achève.

Les archives sont accessibles *via* un outil d'accès : la Wayback machine. Cet outil permet l'interrogation des archives à partir d'une requête par URL, l'affichage chronologique des différentes dates de capture pour chaque URL, la navigation dans les archives en utilisant des liens hypertexte.

Sauvegarde : Après indexation, les données (fichiers ARC et CDX) sont transférées quotidiennement dans SPAR. SPAR est l'infrastructure matérielle du Système d'information numérique de la BnF, chargé de préserver à long terme et de donner accès à tous les documents numériques gérés par la Bibliothèque.

Enfin, deux modules génériques ont été ajoutés pour évoquer les tâches transverses, qui concernent chaque étape du processus de collecte.

Administration et politique de collecte : Cette entité regroupe tous les aspects de définition, de direction et de gestion du processus de collecte : réflexion transverse au sein de la Bibliothèque, communication externe, relations avec les producteurs de sites, gestion des moyens alloués au projet…

Exploitation : Ce module réunit toutes les tâches liées à la maintenance des machines (serveurs de collecte, d'indexation, d'accès) et de l'infrastructure réseau.

La **vérification** des sites proposés à la collecte a été confiée à un agent DLN en fonction des paniers de collecte (un responsable DLN par panier de collecte). C'est lors de cette vérification que l'on peut s'apercevoir des principaux problèmes rencontrés par les sélectionneurs lorsqu'ils spécifient les paramètres techniques.

D'autre part, la vérification suppose au moins un examen rapide du site. C'est ainsi l'occasion d'avoir une connaissance globale et une vision synthétique de toute la collection.

La **planification** a été confiée au responsable de production. Elle nécessite une bonne connaissance du corpus à collecter (taille des sites, fréquence, particularités techniques…) et des contraintes techniques de collecte (temps de collecte, site demandant une surveillance particulière…). Elle est donc liée aux actions de **contrôle de réception** et de **contrôle qualité** – tâches effectuées par les agents chargés de la surveillance.

La **surveillance** était donc de la responsabilité de plusieurs agents DLN, chaque agent avait à sa charge un ensemble à collecter à surveiller (un ensemble à collecter correspondant à un ensemble de sites ayant les mêmes paramètres de fréquence, de profondeur et de budget).

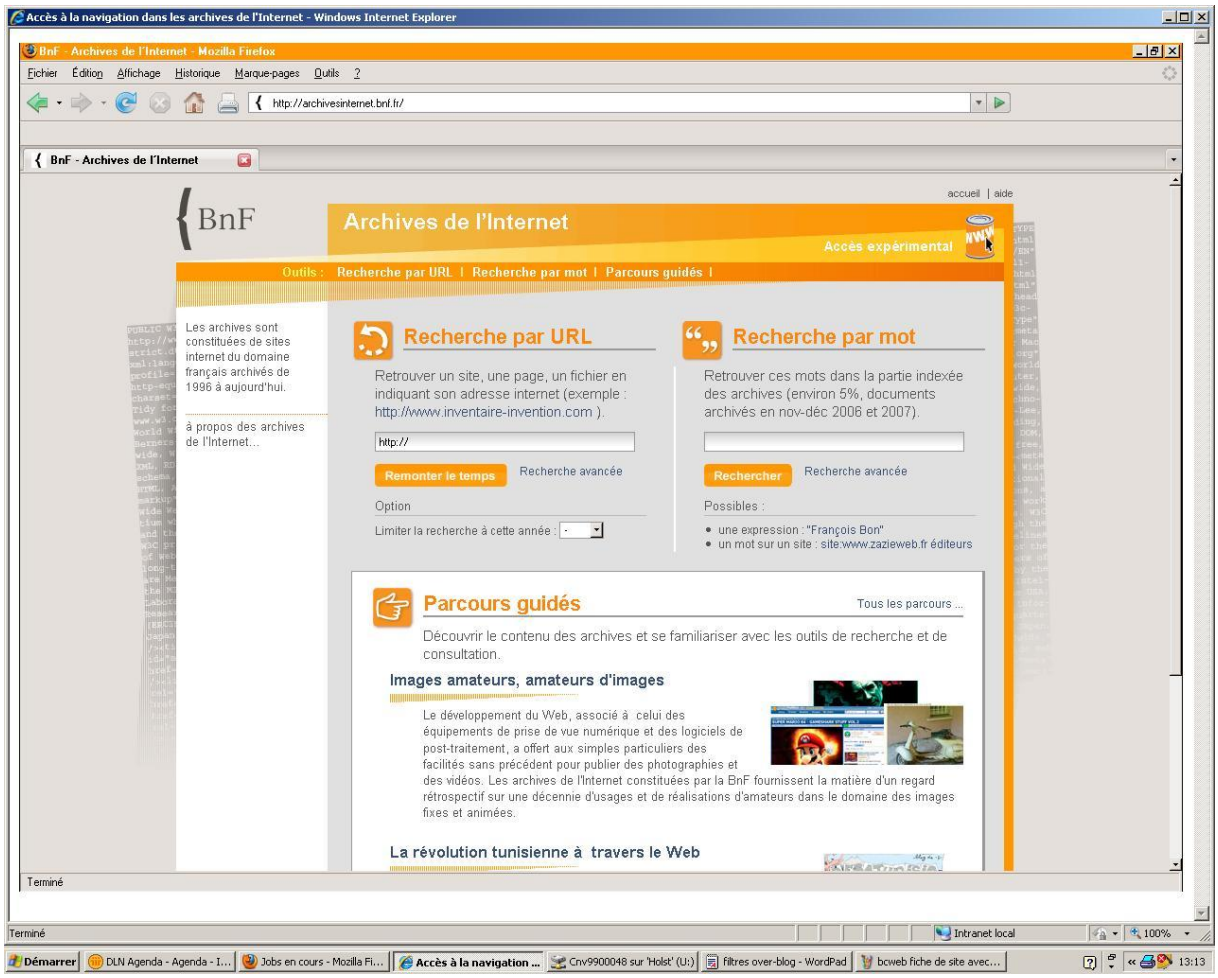
Les tâches de **collecte**, **indexation**, **sauvegarde**, et **exploitation** étaient attribuées aux deux ingénieurs du DSI.

Enfin, **l'administration** était répartie en fonction du niveau :

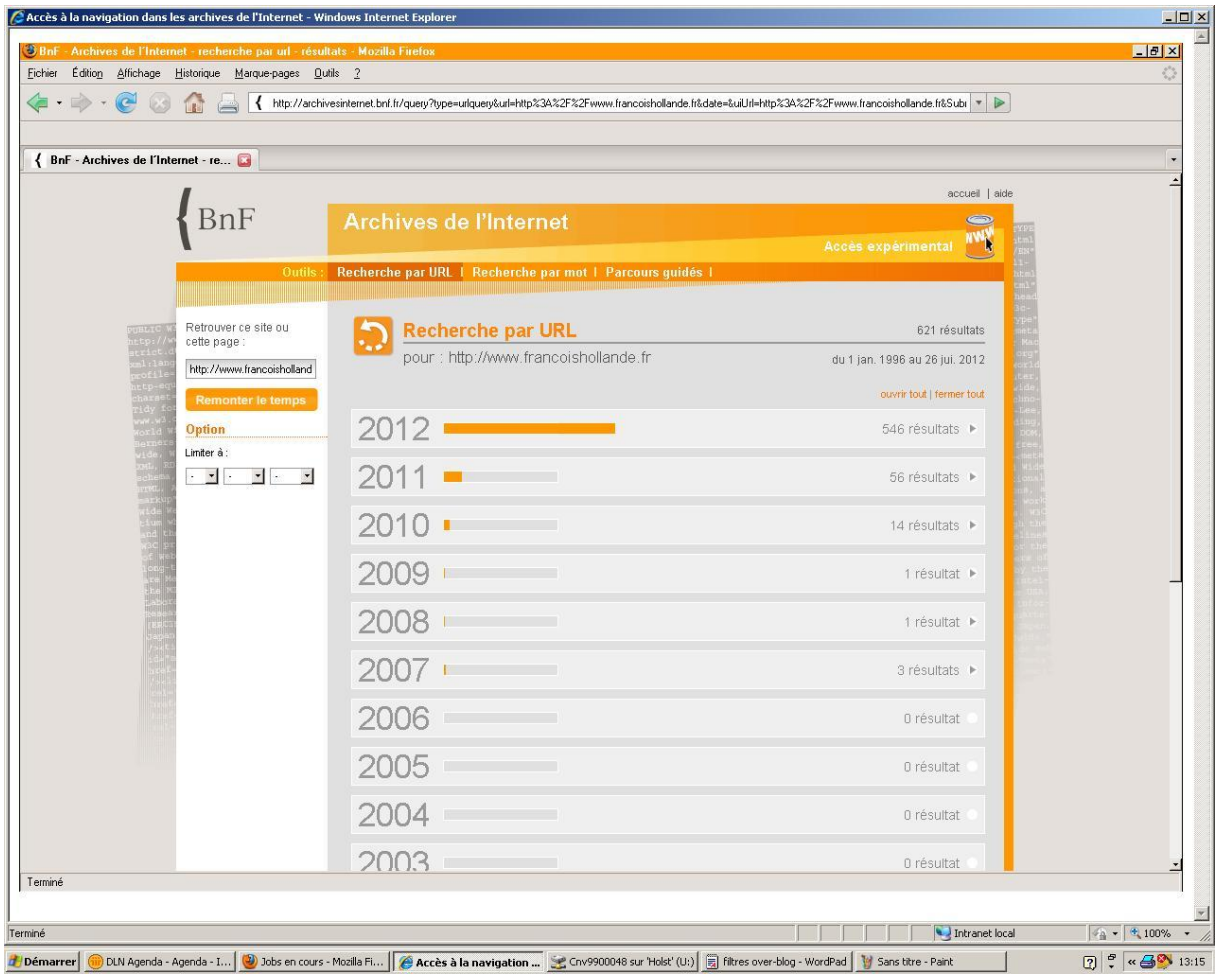
- le niveau politique concerne la définition des objectifs et l'évaluation du projet, les relations stratégiques et institutionnelles avec les partenaires internes et externes de l'équipe (autres départements de la BnF, BDLI), les relations avec les producteurs de sites, la communication externe et interne (conférence de presse…). Il était confié au chef de service DLN.
- la conduite du projet au niveau opérationnel était du ressort des coordonnateurs

Annexe 4 : Interface d'accès, Archives de l'Internet

Page d'accueil :



Suite à une requête, l'interface affiche le nombre de captures disponibles par année :



Après avoir sélectionné une année pour laquelle des captures sont disponibles, il est possible de choisir une date à laquelle le site demandé a été collecté :

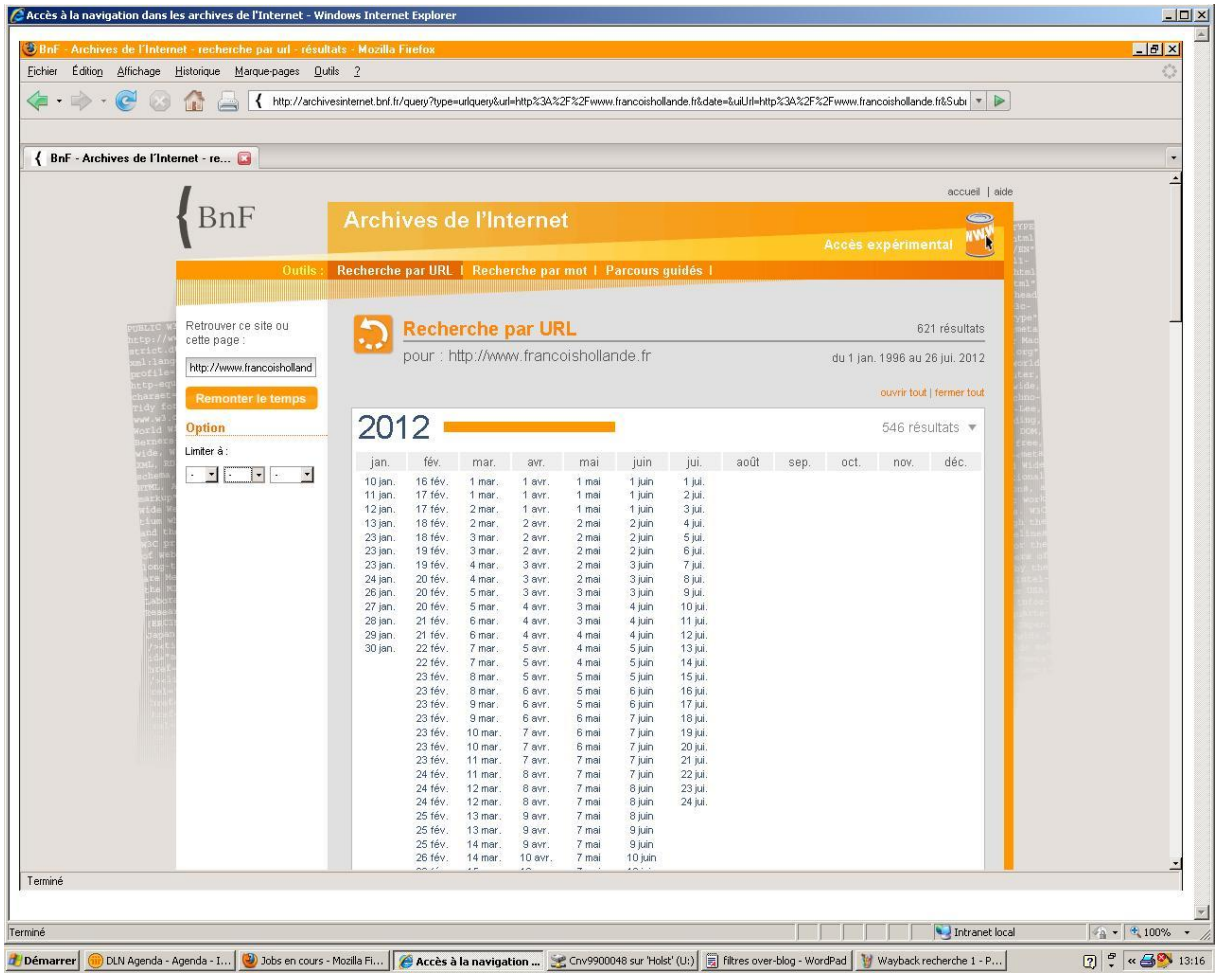


TABLE DES MATIÈRES