Bibliothèque nationale de France

# Collectes électorales 2017 : Bilan

**Bibliothèque nationale de France**

**direction des Services et des réseaux**
département du Dépôt légal
dépôt légal numérique

Version du 11 octobre 2017

# TABLE DES MATIERES

# 1. Documents applicables de référence

*Élections 2017 : Modalités de la collecte électorale 2017,* référence : BnF-ADM-2017-025320-01
*Élections 2017 : Manuel d'utilisation de l'application BCWeb 2017,* référence : BnF-ADM-2017-025323-01
*Elections 2012 : bilan documentaire et organisationnel,* référence : BnF-ADM-2013-023631-01.
*Élections 2010 : bilan de la campagne d'archivage des sites relatifs aux élections régionales*, référence : BnF-ADM-2010-047141-01.
*Élections 2007 : Bilan de l'organisation du projet*, référence : BnF-ADM-2007-031419-01.
*Élections 2007 : Bilan documentaire du projet*, référence : BnF-ADM-2007-034655-01.

Le bilan des collectes électorales 2017 s'appuie également sur les réponses à trois questionnaires envoyés aux acteurs à l'issue du projet, en juillet et août. Ces trois questionnaires de dix questions différaient légèrement en fonction du rôle des acteurs : correspondants dans les établissements partenaires, correspondants à la BnF, et coordonnateurs documentaires. Les taux de retours aux questionnaires sont respectivement de 46% pour les correspondants dans les établissements partenaires (22 réponses sur 48 possibles), 100% pour les coordonnateurs (24 sur 24), 67% pour les correspondants BnF (8 sur 12).
Les réponses étaient anonymes et l'outil utilisé pour la diffusion du questionnaire est Typeform dans sa version gratuite.
Les questions qui différaient concernaient l'organisation du travail propre à chaque établissement et chaque rôle. Le questionnaire avait pour finalité d'avoir des éléments de visibilité sur quatre grands axes : l'organisation et l'assistance dans le cadre du projet côté établissements partenaires (formation, communication, volumes horaires sollicités) ; l'organisation et l'assistance dans le cadre du projet côté établissement BnF ; le recueil de points de vue en matière de périmètre documentaire par les différents acteurs ; et des retours sur l'application et outil de sélection BnF Collectes du Web. Les résultats et réponses sont très précieux pour envisager les choix d'organisation lors des futures collectes électorales, mieux hiérarchiser les documents à sélectionner et les efforts de collecte en découlant, et enfin pour envisager des évolutions sur l'application BCWeb.

# 2. Terminologie

## 2.1. Glossaire

« Archive » d'un site / d'une page : ensemble des captures d'un même site / d'une même page conservées à un même endroit. Exemple : l'archive du site http://www.senat.fr.

BnF - Collecte du Web (BCWeb): outil de saisie des propositions de collecte de sites Web.

Capture : ensemble de fichiers collectés par le robot à une date précise et pour une URL donnée.

Collecte ciblée : collecte spécifiquement lancée sur un nombre de sites ou de parties de sites dont les adresses URL ont été au préalable identifiées par des bibliothécaires. Dans les filières d'entrée du document de dépôt légal du web à la BnF. La collecte ciblée se différencie de la collecte large et du dépôt, qui consiste à archiver un site en travaillant directement avec son producteur.

Collecte courante : type de collecte ciblée dans laquelle les sites identifiés ont vocation à être capturés régulièrement par la BnF.

Correspondant : personne impliquée dans l'action de sélection des sites, comprenant veille, repérage et création des fiches via l'outil BCWeb. Le groupe des correspondants est composé d'agents de la BnF des services DEP (Droit, économie, politique) et PHS (Philosophie, histoire, sciences de l'Homme), ainsi que d'agents issus d'établissements partenaires en charge du dépôt légal imprimeur.

Collecte large : collecte de surface qui permet d'obtenir un échantillon diversifié de pages Web à partir d'adresses URL de départ, obtenues auprès de bureaux d'enregistrements des noms de domaines. Une collecte large peut se « limiter » à un domaine de haut niveau (comme le .fr).

Collecte projet : type de collecte ciblée dans laquelle les sites identifiés par les bibliothécaires ont vocation à être archivés sur une période déterminée. Ils doivent être relatifs à un thème ou à un événement particulier (élection, festival…). À l'issue du projet, certains des sites identifiés peuvent être injectés dans la collecte courante.

Job : Un *job* est une tâche confiée à un robot de collecte, et gérée à partir du logiciel de planification NetarchiveSuite. Chaque *job* regroupe une liste d'adresses de départ ainsi que des paramètres de collecte, notamment la profondeur de collecte, le budget alloué, des règles de comportement spécifiques, des exclusions. Toutes les adresses de départ contenues dans un même *job* sont traitées en parallèle par le robot de collecte. En

pratique, plus il y a d'adresses de départ dans un même *job*, ou plus le budget et la profondeur alloués sont élevés, plus le *job* va durer longtemps - il va également ramener davantage de données.

Robot de capture / de collecte / *crawler* : robot parcourant la Toile, grâce à l'utilisation des liens hypertextes, afin de découvrir et d'archiver des fichiers.

Heritrix : logiciel libre de robot de capture utilisé par la BnF pour l'archivage du web. Sa version 3 est utilisée par la BnF depuis 2017.

## 2.2. Abréviations

BCWeb : BnF - Collecte du Web (outil de saisie et de gestion des collections)
BDLI : Bibliothèque de dépôt légal imprimeur
BnF : Bibliothèque nationale de France
BNUS : Bibliothèque nationale et universitaire de Strasbourg
DDL : Département du Dépôt légal
DLN : Service du Dépôt légal numérique
DLWeb : Dépôt légal du Web
DCO : Direction des Collections
DEP : Département Droit, économie, politique
DSI : Département des Systèmes d'information
DSR : Direction des Services et des réseaux
IIPC : *International Internet Preservation Consortium* (consortium international pour la préservation de l'Internet)
INA : Institut national de l'Audiovisuel
NAS : NetarchiveSuite (outil de gestion des collectes)
PHS : Département Philosophie, histoire, sciences de l'homme
URL : *Uniform Resource Locator*

# 3. Cadre organisationnel et technique

## 3.1. Genèse et calendrier du projet

Le projet de collectes de sites relatifs aux élections présidentielle et législatives de l'année 2017, au titre du dépôt légal numérique, s'inscrit dans la continuité quasi directe des collectes relatives aux élections régionales de décembre 2015 - les deux élections étant espacées d'environ un an. La typologie documentaire et les critères de sélection à l'œuvre depuis les collectes électorales de 2007 ont été repris. Cependant, la mise en place très en amont du calendrier électoral par le ministère de l'intérieur garant de l'organisation des campagnes et des scrutins de 2017 - en comparaison des élections de 2015 situées à une période inhabituelle et intégrant les nouveautés de la Loi NOTRe[1] - comme la dimension médiatique plus conséquente des campagnes de 2017, ont permis d'organiser plus en amont le calendrier du projet en fonction des échéances propres à ces élections. Trois périodes successives de collecte ont ainsi pu être distinguées dès le commencement du projet :

- Les collectes relatives aux primaires EELV, de la droite et du centre, et de la gauche ont été menées du 7 novembre au 1er février, avec les sélections de sites effectuées par une équipe restreinte de correspondants BnF
- Les collectes relatives à l'élection présidentielle (campagne officielle et scrutins) ont été menées du 1er février au 12 juillet, avec les sélections effectuées par une équipe de correspondants BnF, et ponctuellement des équipes dans les établissements partenaires
- Les collectes relatives aux élections législatives ont été menées du 3 avril au 12 juillet, avec les sélections effectuées par une équipe élargie de sélectionneurs BnF et une vingtaine d'équipes de correspondants dans les établissements partenaires

Les équipes de sélections pour ces trois différentes périodes de collecte ont été constituées au fur et à mesure du projet. L'appel à participation auprès des établissements partenaires a été envoyé le 15 février soit 30 jours avant le début des sélections et 45 avant le premier envoi à la collecte de ces sélections. Selon les réponses aux questionnaires, la date d'envoi de l'appel à participation a été jugée satisfaisante par 88% des coordonnateurs documentaires et laissait ainsi le temps de s'organiser en interne.

[1] Loi portant sur la Nouvelle Organisation Territoriale de la République. Pour plus d'informations http://www.gouvernement.fr/action/la-reforme-territoriale ou http://archivesinternet.bnf.fr/20170301140109/http://www.gouvernement.fr/action/la-reforme-territoriale (dans les archives de l'internet de la BnF)

**Calendrier du projet**[2]

- Octobre 2016 : début des sélections pour les collectes relatives aux primaires (première collecte de sites le 7 novembre 2016)
- Janvier 2017 : constitution des équipes de correspondants BnF pour la collecte des sites relatifs à l'élection présidentielle
- 31 janvier : première réunion du comité de sélection BnF
- 1er février 2017 : dernière collecte relative aux primaires, premières collectes hebdomadaires et mensuelles relatives à la présidentielle
- 15 février : appel à participation auprès des établissements partenaires pour la sélection des sites relatifs aux élections législatives
- 20 mars : envoi du manuel de sélection et du calendrier du projet et des collectes aux établissements partenaires ; ouverture de l'outil de sélection BnF-Collectes du web aux correspondants externes
- 22 mars : lancement des collectes pluriquotidiennes spécifiques aux réseaux sociaux
- 30 mars : deuxième réunion du comité de sélection BnF
- 3 avril : premières collectes mensuelle et hebdomadaire ouvertes aux sites relatifs aux législatives
- 22 et 23 avril : premier tour de la présidentielle
- 2 mai : troisième réunion du comité de sélection BnF
- 6 et 7 mai : second tour de la présidentielle
- Lundi 29 mai : dernière collecte mensuelle avant le premier tour des législatives
- 11 juin : premier tour des législatives (3 juin pour la circonscription de l'étranger)
- 18 juin : deuxième tour des législatives
- 29 juin : quatrième réunion de bilan du comité de sélection BnF
- 30 juin : dernier jour des collectes pluriquotidiennes dédiées aux réseaux sociaux
- 11 juillet : envoi de questionnaires aux participants pour un recueil d'avis sur l'organisation du projet
- 12 juillet : dernière collecte rassemblant tous les sites sélectionnés et encore actifs en périodicité hebdomadaire et mensuelle
- Août 2017 : rédaction du bilan du projet des collectes électorales et publication des listes de sites sélectionnés sur data.gouv.fr

## 3.2. Critères de sélection et cadre documentaire

Dans une perspective de continuité des collections, la plupart des choix documentaires, à l'œuvre depuis 2007 et constitutifs de la forme de la collecte, ont été reconduits.

### *3.2.1. Critères de sélection des sites*

Les critères de sélection « discriminants » sont identiques à ceux des projets de 2010, 2012 et de 2015. Communs aux élections présidentielles et législatives, ces critères de sélection servent à la bonne identification et sélection des sites par les correspondants.

Soit :

- Des critères généraux du périmètre de la collecte de sites de dépôt légal à la BnF

Les publications en ligne sélectionnées doivent appartenir au « domaine français » couvert par la mission de dépôt légal de l'internet de la BnF, juridiquement défini par le code du patrimoine, aux articles (art. L131-2, L132-2, L132-2-1 et R132-23-1). « les signes, signaux, écrits, images, sons ou messages de toute nature faisant l'objet d'une communication au public par voie électronique ». Les documents entrant dans cette définition doivent en outre, pour pouvoir être collectés, soit :

- appartenir à l'extension .fr
- être produits par un auteur domicilié en France
- être édités par un auteur résidant en France
- être produits par des moyens français

Enfin, ils ne doivent pas appartenir au périmètre couvert par le dépôt légal numérique de l'INA, concernant les sites des radios et télévisions françaises.

- Un critère de contenu

Les sites ou parties de sites retenus doivent êtres relatifs aux élections présidentielle et/ou législatives de 2017.
Ce premier critère de sélection se fait donc indépendamment de la forme du site diffusé.
Pour des sites généralistes traitant des élections parmi d'autre sujets, la sélection devait s'appliquer à la partie de site correspondante (rubrique, fils d'actualité, ensemble d'articles).

[2] Le calendrier complet des dates de collecte est à retrouver en annexe 3

- Un critère d'actualité

Les sites sélectionnés doivent être mis à jour avec régularité durant la période de la campagne

- Un critère de représentativité

A l'échelle de l'ensemble de la sélection, un équilibre idéologique doit être maintenu autant que possible entre les différents courants de pensée qui s'expriment en ligne. La diversité des débats et toutes les grandes tendances se manifestant en ligne doivent être couverts dans leur juste proportion, dans un but de représentativité du pluralisme politique et idéologique disponible en ligne.

- Un critère de navigabilité

Le site ou les parties de sites les plus significatifs doivent être sélectionnés, afin de ne pas collecter des pages et documents non pertinents par rapport au sujet. Malgré tout, l'unité documentaire de référence demeure le site : la sélection « à la pièce » de l'adresse d'un article, d'un seul élément d'un site ou d'une page isolée était à éviter.

### *3.2.2. Typologie, thèmes et mots-clés*

La typologie de sélection par émetteur, utilisée à chaque collecte électorale depuis 2007, a été reconduite pour la collecte du web électoral de 2017. Elle permet d'assurer la continuité du contenu des collections dans le temps, sans exclure les nouveaux formats de sites s'y intégrant de fait, car nécessairement émis par une personne ou une organisation.

Deux grands types de contenus sont couverts :
- un contenu émanant de la classe politique, avec une volonté de pluralisme et de représentativité de l'ensemble des courants politiques.
- un contenu émanant du corps social, dans une logique de représentativité et d'échantillonnage induits par le grand volume de données disponibles.

La typologie s'organise selon le cadre de classement suivant :

**0- Sites officiels et institutionnels :** institutions qui veillent à l'organisation des élections et au respect du cadre juridique

**1- Les candidats et leurs organisations**
1.1 Sites des candidats en campagne
1.2 Sites des formations politiques
1.3 Autres organisations de soutien

**2- Regards et opinions sur la campagne**
2.1 Annuaires, observatoires et analyses
2.2 Médias traditionnels
2.3 Associations, syndicats et autres organisations
2.4 Expressions individuelles et communautaires sur l'Internet

### *3.2.3. Calendrier, paramétrages et volumétrie cible es collectes*

Les correspondants en charge de la sélection des sites internet ont dû choisir entre différents paramètres de collecte (périodicité, budget en nombre de page à collecter, profondeur pour déterminer quelles parties du site sélectionner) pour chacun des sites sélectionnés. Après la veille et le choix du site à collecter, le rôle majeur du sélectionneur est de faire son choix parmi ces paramètres afin d'établir une stratégie de sélection et obtenir la meilleure couverture temporelle et documentaire possible de son corpus de sites.
Le calendrier des **périodicités**, premier des trois paramètres, est en étroite dépendance avec le calendrier des campagnes et scrutins électoraux[3]. Il comportait trois grandes périodicités :
- **Une collecte pluriquotidienne** : Deux collectes par jour à 11h et 23h ont eu lieu sans interruption du 22 mars au 30 juin. Elle était dédiée aux comptes et mots-dièses (hashtag[4]) des réseaux sociaux Twitter et Instagram.
- **Une collecte hebdomadaire** a eu lieu tous les lundis à 14h du 6 février au 10 juillet sans interruption. Elle était dédiée aux sites classiques dont les contenus se renouvellent à un rythme élevé et nécessitant une collecte plus régulière que mensuelle. Seule la profondeur page+2 était disponible pour cette collecte.
- **Une collecte dite « mensuelle »** a eu lieu du 1er janvier au 12 juillet. Elle a eu lieu chaque 1er du mois de janvier à avril, puis à des dates plus resserrées (environ toutes les trois semaines) et adaptées aux moments

[3] Le calendrier complet des dates de collecte est à retrouver en annexe 2
[4] Le hashtag a été installé en 2009 lorsque Twitter a transformé sur sa plateforme toute expression précédée d'un « hash » (#) en lien hypertexte vers une page compilant tous les messages comprenant cette expression.

clés des campagnes et scrutins. Il s'agit de la fréquence principale, car la plus conséquente en nombre de sites, permettant de sélectionner en profondeur hôte, domaine et page +2, avec un budget moyen.

Chacune de ces collectes était associée à un **budget** en nombre de pages URL à collecter par site**.** Les collectes hebdomadaires et mensuelles permettaient l'attribution d'un budget de type « moyen », soit une volumétrie maximale cible de 90 000 éléments de page ou URL par site. Les comptes et mots-dièses en collecte pluriquotidienne étaient collectés avec un budget petit[5].

Le dernier paramètre qui entrait en compte pour moduler la sélection était la **profondeur**, qui permettait de prendre tout ou une partie de site. Quatre profondeur étaient disponibles : une profondeur domaine (correspondant à la collecte de la totalité du site selon la syntaxe domaine.extension) ; une profondeur hôte (permettant de collecter une certaine partie du site hébergée de manière individuelle, discernable à partir de la syntaxe type hôte.domaine.extension) ; une profondeur page +2clics, permettant de collecter une page simple (par exemple une page d'accueil) ainsi que tous les contenus se trouvant à 2 clics sur les liens de cette page ; et enfin une profondeur websocial, consacrée à la collecte des réseaux sociaux Twitter et Instagram.

## 3.3. Les moyens humains

Lors de ses différentes étapes, les collectes électorales ont impliqué la participation, bien sûr différenciée en matière de temps de travail, de 111 personnes, pour des activités aussi diverses que la mise en place des outils de collecte, la coordination du projet, la sélection des sites à collecter…

Sur le strict point de vue du repérage et du choix des sites, les équipes de sélections se sont répartis entre une équipe de sélection à la BnF qui a rassemblé, à différents moments du projet, jusqu'à 18 personnes. Les équipes de sélectionneurs BDLI ont rassemblé 70 participants pour l'ensemble des collectes législatives.

### *3.3.1. L'équipe de sélection BnF*

Le terme d'équipe de sélection BnF regroupe les personnes impliquées dans la sélection au sens large : de la coordination documentaire générale des collectes à l'action de sélection des sites en elle-même (impliquant veille, repérage et création des fiches).

**Quatre agents** ont été en charge de la coordination technique et documentaire : 2 personnes du DDL (dont un coordonnateur et une gestionnaire), une personne du département PHS, et deux personnes, successivement, du département DEP. La coordination technique recouvrait l'ensemble des tâches de conception et conduite du projet : préparation des critères de sélection et du calendrier, formation et assistance des sélectionneurs sur l'application de sélection BnF – Collectes du web, préparation et animation des comités de sélection ; tandis que la coordination documentaire recouvrait l'organisation et l'animation des équipes de sélections, le suivi documentaire des propositions de collecte, et la normalisation de la description documentaire (mots-clés).

**12 agents de la direction des Collections** composaient le groupe des correspondants : trois personnes du service **Droit science politique publications officielles**, une personne du service Presse et deux personnes du service Économie de DEP; trois personnes du service Histoire et trois personnes des services sciences sociales de PHS. L'équipe a été renforcée de trois personnes pour la sélection des sites de la campagne des législatives dans la région Provence-Alpes-Côte-D'azur. Pour la collecte dédiée aux vidéos et contenus audiovisuels spécifiques, un renfort a été apporté en cours de projet par trois correspondants du département de l'Audiovisuel.

Le travail de sélection a été réparti sur trois grands types de collectes :

- La collecte relative à l'élection présidentielle
- La collecte relative aux élections législatives dans les circonscriptions d'Ile-de-France
- La collecte relative aux élections législatives dans les circonscriptions de Mayotte, Wallis et Futuna, de Polynésie Française, de Saint Pierre et Miquelon, des français de l'étranger (couverture BnF prévue au départ) et de la Corse, Nouvelle-Calédonie et de Provence-Alpes-Côte d'Azur. Cette dernière région, dont la couverture par la BnF n'avait pas été prévue, a représenté un surcroît de travail difficile à anticiper, qui a nécessité le renfort de trois sélectionneurs. Pour les prochains projets, une équipe plus nombreuse de correspondants est indispensable, au moins durant le temps des élections législatives.

Au sein de chaque collecte, les sélectionneurs se sont généralement répartis par thématique du cadre de classement, avec cinq sélectionneurs pour le 0 - Sites officiels et institutionnels / 1.1 Sites des candidats en campagne / 1.2 Sites des formations politiques, trois sélectionneurs pour le 2.1 Annuaires, observatoires et analyses / 2.2 Médias traditionnels / 2.3 Associations, syndicats et autres organisations, et quatre sélectionneurs

[5] Plus d'informations sur la collecte des comptes et mots-dièse Twitter et Instagram est disponible dans la partie3.4.3

pour le 1.3 Autres organisations de soutien / 2.4 Expressions individuelles et communautaires sur l'Internet . Ces quatre derniers ont également mené la réflexion pour circonscrire le périmètre du 2.4 potentiellement infini.

Enfin, comme pour chaque projet, les correspondants BnF se sont réunis lors de quatre comités de sélections. Ces réunions régulières, organisées à trois reprises au cours du projet, avaient pour objectifs de permettre l'échange sur les grandes tendances de la campagne en ligne et de réajuster dans le détail la collecte, la sélection et la description documentaire. Elles ont notamment conduit aux ajustements suivants :

- Adaptation volumétrique : ampleur de la collecte des sites de la typologie 0 dans la collecte de niveau national; ampleur de la couverture de la présence numérique d'un candidat, avec la multiplication de comptes sur les réseaux sociaux pour une même personne ; ampleur de la couverture des candidats des listes électorales (choix de la sélection des premiers noms des listes électorales)
- Ajustement de la description documentaire, dont la validation d'une liste des mots-clés pour les partis et les mouvements de 2017
- Adaptation de la juste proportion de sélections pour les expressions individuelles et communautaires sur l'internet, car la cible documentaire potentiellement très vaste.
- Discussions sur les tendances de publications politiques en ligne en 2017. Adaptation par rapport aux propositions externes de sites à collecter proposés via Twitter, et par des particuliers via la boîte générique du service du DLN ou le compte Twitter @Dlwebbnf ou par des collègues hors du projet.
- Organisation du contrôle qualité dans les archives de l'internet sur les sites sélectionnés et collectés tout le long de la période
- Discussion sur la bonne organisation des collectes et les règles d'harmonisation dans l'application BnF Collectes du web

### *3.3.2. Les équipes de sélections dans les établissements partenaires*

Les sélectionneurs des BDLI ont eu un rôle de sélection local à l'échelle de l'ancienne région à laquelle l'établissement appartient. Ils ont parfois signalé des sites d'envergure nationale à intégrer dans la collecte de niveau national.

**23 établissements** de dépôt légal imprimeur ont donné suite à l'appel à participation, ce qui a conduit 70 personnes à participer à l'activité de sélection. Le réseau des bibliothèques de dépôt légal imprimeur étant structuré sur le modèle des anciennes régions, chaque établissement a couvert les campagnes relatives aux circonscriptions législatives dans le périmètre des départements dont il a la charge au titre du dépôt légal imprimeur.

La couverture était la suivante :

- Guadeloupe : Archives départementales de la Guadeloupe
- Guyane : Archives départementales de la Guyane
- La Réunion : Archives départementales de La Réunion
- Martinique : Archives départementales de la Martinique
- Région Auvergne-Rhône-Alpes : Bibliothèque municipale de Lyon, Bibliothèque municipale de Clermont-Ferrand
- Région Bourgogne-Franche-Comté : Bibliothèque municipale de Dijon, Bibliothèque municipale de Besançon
- Région Bretagne : Bibliothèque municipale de Rennes
- Région Centre-Val de Loire : Bibliothèque municipale d'Orléans
- Région Corse : Bibliothèque nationale de France
- Région Grand-Est : Bibliothèque municipale de Nancy, Bibliothèque nationale et universitaire de Strasbourg, Bibliothèque de Châlons-en-Champagne
- Région Hauts-de-France : Bibliothèque municipale de Lille, Bibliothèque municipale d'Amiens
- Région Île-de-France : Bibliothèque nationale de France
- Région Normandie : Rouen nouvelles bibliothèques, Bibliothèque municipale de Caen
- Région Nouvelle-Aquitaine : Bibliothèque municipale de Bordeaux, Bibliothèque municipale de Poitiers, Bibliothèque francophone multimédia de Limoges
- Région Occitanie : Bibliothèque municipale de Toulouse, Médiathèque Montpellier Méditerranée Métropole
- Région Pays-de-la-Loire : Bibliothèque municipale d'Angers
- Région Provence-Alpes-Côte d'Azur : Bibliothèque nationale de France

Les sélectionneurs étaient principalement des agents des services en charge du dépôt légal imprimeur ou du service Patrimoine des établissements partenaires.

Le service du dépôt légal numérique de la BnF n'a pas proposé de journée de formation spécifiquement dédiée pour l'organisation de la sélection au sein de chaque établissement. En substitution, elle a élaboré deux documents de référence (un document cadre sur l'ensemble du projet et un manuel du sélectionneur version 2017 en vue de la manipulation de l'outil BnF – collecte du web). Des formations par téléphone ont également été proposées. Une seule équipe a demandé à bénéficier de cette séance de formation. Au cours de la période de sélection, des messages généraux sur le déroulé du projet et sur les questions fréquentes ont été envoyés à intervalles réguliers. En outre un interlocuteur référent au DLN était désigné pour chaque collecte. Il a été amené à échanger régulièrement par téléphone ou courriel avec les différents correspondants.

Sur cette problématique de la formation des acteurs du projet, les retours des questionnaires sont plutôt positifs. Le support de formation a été jugé comme donnant suffisamment d'informations par 100% des répondants coordonnateurs documentaires, 90% des sélectionneurs. Néanmoins, 36% ont jugés qu'il était nécessaire de l'accompagner par davantage de points d'étapes de communication en cours du projet (dont des documents regroupant les questions fréquemment posées) et surtout 40% estiment qu'il est nécessaire de proposer parallèlement une formation spécifique à cette activité, demandé explicitement dans trois réponses au questionnaire des coordonnateurs.
Ainsi, il faudrait envisager d'organiser pour les futurs projets d'envergure des journées de formations générales, sur le modèle de la journée de formation organisée à la BnF pour les collectes électorales de 2012, ou sur un modèle de formation particulière. Ces formations sont toujours proposées à chaque nouveau correspondant intégrant le réseau de sélection interne de la BnF. Il faudrait également à l'avenir intensifier la communication interne au cours du projet.

### *3.3.3. L'équipe DLweb*

Le terme d'équipe DLweb désigne ici les agents du service du dépôt légal numérique, au sein du département du dépôt légal, et 4 agents du département des systèmes d'information mobilisés ponctuellement sur ce projet.

Les agents du département des systèmes d'information ont maintenu en service les outils de collecte et effectué des tests techniques en février en vue de la collecte des réseaux sociaux

Comme lors des précédents projets, les agents du DLN étaient eux chargés de la gestion technique des collectes. La gestion du planning consistait à veiller à la bonne organisation des collectes afin qu'elles puissent être réalisées selon la planification établie en accord avec le comité de sélection et suivant le calendrier électoral.

En outre, les sept agents du DLN avaient des tâches diverses. Chacun était responsable technique de plusieurs collectes régionales (de trois à cinq selon leur ampleur), ce qui regroupait l'ensemble des tâches de coordonnateur technique. Tout d'abord, un rôle d'interlocuteur technique pour les correspondants des collectes : assistance sur les transferts de fiches, sur les paramètres optimaux pour la collecte de sites particuliers, sur les questions de périmètre documentaires ou de nature de documents à collecter…
Ensuite, une tâche de vérification et de surveillance leur était également allouée. La saisie de toutes les URL entrées dans BCWeb, avant collecte, était vérifiée, du point de vue des paramètres techniques, du périmètre documentaire et de la cohérence avec les autres sélections (doublons). La plupart des corrections a porté sur les paramètres de profondeur du site (inversion hôte/domaine, pages sélectionnés à l'unité…) ou sur la syntaxe des URL. La surveillance est une action qui vient une fois les collectes lancées, pour s'assurer de leur bon déroulement. Cela consiste à éviter que le robot de collecte ne s'enferre dans des pièges (calendriers, javascript), et à contrôler que la collecte se déroule normalement (prise en compte des blocages, des redirections, etc.).

Un contrôle de réception, à la fois pendant et après les collectes, a été mené sur un échantillon de 10% de la collecte. Pour les sites dont la mauvaise collecte était due à un blocage par le producteur, volontaire ou involontaire, du logiciel de collecte de la BnF (repéré au moment de la surveillance ou du contrôle), une vingtaine de prises de contacts avec des producteurs ont été effectuées, qui ont abouti la moitié du temps à un déblocage.

## 3.4. Moyens techniques

### *3.4.1. L'application BnF – Collectes du web*

L'application de collecte **BCWeb** mise en place en 2012 et utilisée depuis pour l'ensemble des sélections en collecte courante, a été utilisée par l'ensemble des sélectionneurs de la collecte projet.
Des collectes par anciennes régions ont été créées dans l'application, sur le format « Collecte Projet / élections 2017 / nom de l'ancienne région » afin de faciliter la répartition des sélections pour les groupes de sélectionneur par établissement. Chaque sélectionneur était uniquement habilité à créer et modifier des fiches dans la collecte de

sa région, mais avait la possibilité de faire des recherches sur l'ensemble des fiches de la base et de consulter l'ensemble des collectes.
Selon les différents retours, l'application est relativement simple d'utilisation et ergonomique. La principale difficulté d'usage d'origine technique demeure le blocage de certaines URL valides en URL invalides, surtout des sites en HTTPS, obligeant à enlever le « s » final et passer outre la redirection proposée.
Dans le cadre du questionnaire, plusieurs questions portaient sur l'application, dans l'optique de pouvoir faire évoluer à moyen terme certaines de ces fonctionnalités. De prime abord, l'application est perçue de manière générale comme fonctionnelle et ergonomique. Cependant la question « Quelle fonctionnalité ou absence de fonctionnalité est selon vous la plus contraignante pour la conduite du travail de sélection ?» a permis de relever quelques difficultés techniques inhérentes à l'application et au paramétrage des collectes. La profondeur page+2 et la pertinence de son utilisation est parfois jugée comme difficile à déduire. La gestion de l'outil de repérage des doublons d'adresses déjà sélectionnées dans la base pourrait être plus efficace selon certains cas, car elles ne permettaient pas, en partie, de repérer par exemple si un compte Twitter était bien sélectionné ou non (le dédoublonnage et la présentation de la liste se faisant d'abord au niveau domaine). Des fonctionnalités de thésaurus pour les mots-clés et de visualisation autre que par liste (tableau synoptique, carte heuristique à partir des mots-clés, graphiques) ont été évoquées comme fonctionnalités suggérées.
La fonctionnalité de repérage des doublons dans la base est par ailleurs jugée comme la fonctionnalité la plus utile de l'application, à partir des réponses à la question « Quelle fonctionnalité ou absence de fonctionnalité est selon vous la plus utile pour la conduite du travail de sélection ?». Parmi beaucoup d'autres, la fonctionnalité d'extraction des listes sous forme de tableurs a également été relevée comme utile.
Les réponses à la question sur l'amélioration et nouvelles fonctionnalités souhaitées permettront-elles de nourrir le projet d'évolution de l'application en 2018.

A partir des commandes de collecte effectuées dans BCWeb, et après versement dans cette application des adresses sélectionnées, l'outil de planification NetarchiveSuite permettait de planifier les collectes et de mener les tâches de surveillance et certaines de contrôle qualité. Il s'agissait d'une nouvelle version (la 5ème) de ce logiciel de planification, utilisé depuis mars 2017 à la suite du passage de la BnF de la version 1 à la version 3 de son logiciel de collecte Heritrix.

### *3.4.2. Heritrix 3 et cas particuliers de collectes*

La principale nouveauté technique à l'œuvre lors des collectes électorales réside dans l'utilisation d'une nouvelle version du logiciel de collecte utilisé par la BnF à partir du mois de mars 2017. Le projet de passage à NetarchiveSuite 5 et Heritrix 3 a été mené par l'équipe DLweb entre juillet 2016 et mars 2017, avec pour objectif d'avoir ce nouvel outil prêt pour les collectes électorales de la présidentielle et des législatives. Heritrix 3 est également désormais utilisé pour l'ensemble des collectes courantes et projets.
Passer à NetarchiveSuite 5 et Heritrix 3 était une nécessité pour avoir des outils adaptés à la collecte des contenus et aux protocoles utilisés sur le web aujourd'hui. Les collectes sont de fait de meilleure qualité et s'étendent désormais aux sites utilisant le protocole HTTPS. Le bruit et les erreurs 404 provoquées par le code Javascript ont également diminué grâce au nouvel extracteur Javascript. Ainsi les collectes sont moins volumineuses en stockage. La récupération des feuilles de style des sites et des images est aussi plus efficace.
Le passage à Heritrix 3 a également permis de tester et de redéfinir le mode de collecte des comptes Twitter et d'avoir un plafond supérieur (environ 3000 désormais) en nombre de comptes capturés quotidiennement et en parallèle.

Certains contenus publiés sur des grandes plateformes type réseau social ou vidéo ont dû faire l'objet d'aménagements particuliers pour leur bonne collecte. Pour une plateforme en particulier, la collecte s'est avérée impossible d'un point de vue technique et a dû être sciemment écartée de la politique documentaire.

Pour **Facebook** en effet, la mise en place depuis 2015 d'un code de sécurité, de type captcha, pour accéder aux pages publiques sans être connecté sur le réseau, a empêché la récupération de ces pages avec le logiciel de collecte. Ce contrôle de sécurité sert à discerner les utilisateurs humains des procédures automatisées. Ces dernières peuvent être de toute sorte (robot d'indexation, de collecte, de diffusion de spam) et incluent donc le logiciel de collecte de la BnF. Aucune solution de contournement de ce code n'a pu être trouvée à l'heure actuelle, obligeant à laisser de côté ce contenu important, au vu de son audience et de sa volatilité.

Pour **Twitter,** la collecte s'est améliorée par rapport à 2015 tout en restant relativement incomplète. Jusqu'à 3000 comptes et mots-dièses ont pu être collectés quotidiennement. Ces collectes quotidiennes permettaient de collecter les 40 tweets et retweets les plus récents par jour sur chaque compte. Les photos et les liens (liens hypertextes, mots-dièse, autres comptes cités dans les tweets) ont été bien récupérés ainsi que le contexte de la plateforme (disposition en flux, maquette originale de Twitter.com) ce que ne permet pas de faire la récupération par API, l'autre grande alternative de récupération du contenu. Seuls le contenu vidéo et les images de prévisualisation d'articles intégrés à un tweet sont manquants.

Pour **Instagram**, réseau social de plus en plus utilisé et appartenant à Facebook, la récupération des pages a été plutôt bonne et la sélection conseillée : la collecte quotidienne permettait de récupérer les dix images et publications les plus récentes de chaque compte.

Enfin pour la plateforme de vidéo **Youtube,** la récupération de vidéos a dû faire l'objet d'une collecte distincte, effectuée par un prestataire externe. En effet, le circuit d'archivage web du dépôt légal numérique ne permet pas à ce jour de collecter des contenus vidéo sur Youtube. Il avait donc été demandé aux sélectionneurs de ne pas viser des contenus vidéo. La prestation externe a elle concerné la collecte de 27 chaînes : les chaînes des candidats aux primaires et à l'élection présidentielle ; les chaînes de campagne de ces mêmes candidats ; les chaînes de leur formation politique. Les 9 235 vidéos (pour un volume des données collecté s'élevant à près d'un téraoctet) seront disponibles au sein de l'application de consultation dans un second temps.

## 4. Synthèse des résultats des collectes

La veille et le repérage de sites à collecter ont représenté une part importante de la mission des correspondants. La veille a commencé dès le mois d'octobre 2016, dans le cadre des premières collectes relatives aux primaires. Dès le mois de février 2017, les correspondants de la BnF ont démarré la veille et la sélection de sites relatifs à l'élection présidentielle. Les correspondants des BDLI ont pu commencer à sélectionner des sites relatifs aux élections législatives dès le mois d'avril. Ils ont été rejoints par les correspondants BnF qui ont également assuré une veille et un repérage de sites relatifs aux élections législatives pour certaines circonscriptions non couvertes par les BDLI ainsi que pour la région Ile-de-France. La veille et le repérage de sites web a duré jusqu'à la fin du mois de juin 2017, peu avant les dernières collectes post-élections.

Comme pour les précédentes éditions de collectes du web électoral, les correspondants ont utilisé tout l'éventail des outils de veille disponibles en ligne. Cela inclut :

- les moteurs de recherche généralistes tels que **Google** (http://www.google.fr) ou **DuckDuckGo** (https://duckduckgo.com/) et leurs moteurs de recherche dédiés à l'actualité, à l'image de **Google Actualités** (https://news.google.fr*)* qui compile des sources d'informations du monde entier.
- les annuaires des principaux partis pour trouver les fédérations locales, les sections et sous-sections, les candidats. Par exemple :
    - http://www.parti-socialiste.fr/
    - https://www.republicains.fr/federation
    - http://www.frontnational.com/federations-front-national/
    - https://legislatives2017.lafranceinsoumise.fr/
    - https://legislatives.upr.fr/
    - https://en-marche.fr/comites
    - http://eelv.fr/contact-en-region/

- les portails politiques tels que **Politiquemania** (http://www.politiquemania.com)
- les sites et/ou blogs qui analysent l'actualité politique
- l'encyclopédie collaborative **Wikipédia**, qui propose de nombreux contenus sur les élections et les candidats, enrichis de liens externes
- les réseaux sociaux, utilisés à la fois comme éléments de veille et comme contenus directement pertinents à collecter : Twitter, Instagram, Pinterest et, seulement pour la veille, Facebook (cf. 3.4.2). Les moteurs de recherche de Twitter et Facebook ont aussi pu être utilisés pour trouver les sites et comptes de candidats, parfois difficiles à identifier sur le web.

Chaque correspondant était libre d'utiliser ses sources et de développer ses propres méthodes de travail. Si certains correspondants ont choisi de lisser leur volume d'heure de travail (comprenant veille, sélection, relecture) sur toute la durée du projet, la majorité d'entre eux (88% des correspondants de la BnF, 68% des correspondants et 52% des coordonnateurs des BDLI) l'ont concentré autour des périodes de scrutins.

Le questionnaire envoyé au mois de juillet à l'ensemble des correspondants du projet permet de dégager des tendances communes : la veille et le repérage des sites sont les activités qui les ont le plus occupés, aussi bien pour les élections législatives que pour l'élection présidentielle. La majorité des correspondants de la BnF estiment que les collectes des législatives les ont davantage mobilisés que la collecte de l'élection présidentielle. Ainsi 63% des correspondants de la BnF estiment que leur volume d'heure de travail était supérieur pour la collecte des

législatives. La seconde activité la plus importante pour la grande majorité des correspondants était la création de fiches dans BCWeb (saisie des URL, choix des paramètres).

La typologie des sites collectés balaye un spectre large. Pour l'ensemble des correspondants, les types de sites les plus importants à sélectionner ont été les sites officiels de partis et/ou de candidats ainsi que les comptes/hashtag Twitter publics; Venaient ensuite les sites de presse et les pages Facebook publiques, bien qu'il n'ait pas été possible de collecter ces dernières cette année.

## 4.1. Articulation avec les collectes courantes et réinvestissement des sélections antérieures

Une bonne couverture des sites relatifs aux élections nécessite la mise en place d'un projet spécifique de sélection partagée, à la fois pour faire face et trier la quantité de sites disponibles et de par le calendrier électoral, contraint et immuable. La sélection se place dans la continuité avec les collectes courantes du dépôt légal numérique. Des sites compilés dans d'autres collectes, faisant l'objet de capture tout au long de l'année et parfois à un rythme quotidien, ont nécessairement traité des élections de 2017. Ils pouvaient se trouver dans la collecte Actualités, qui regroupe 166 sites d'actualité en ligne, dont la page d'accueil est collectée quotidiennement en profondeur page+1clic. Cela a permis la collecte de tout article concernant les élections publié en une. Des parties évoquant spécifiquement les élections sur des sites d'actualité volumineux ont néanmoins pu faire l'objet de sélection particulière. Enfin, les titres de presse quotidienne régionale obtenus par le biais de la collecte Presse Payante et la collecte du département DEP, avec ses thèmes « sciences politiques » et « presse » étaient également de bons compléments à la collecte projet.

La sélection de ces sites déjà désignés aurait généralement constitué des doublons. Il a fallu néanmoins, pour certaines publications pertinentes, mais à forte volumétrie ou collectées à des fréquences plus extensives, effectuer une nouvelle sélection avec d'autres paramètres de fréquence ou de profondeur que ceux employés dans les collecte courantes. Une capture au mois de décembre, ou le passage du robot de collecte sur la partie de site la plus pertinente dans le cadre des élections régionales, était assurée par ce biais.

Les sélectionneurs participant à la collecte élections 2017 ont également pu bénéficier de la liste des sites collectés par le passé dans le cadre des autres collectes projet élections depuis l'ouverture de BCWeb en 2012. Ainsi, lorsqu'un site était encore pertinent en 2017, et que son URL n'avait pas changé, chaque sélectionneur avait la possibilité de récupérer la fiche dans la collecte d'origine et de la réactiver pour les besoins de sa sélection en 2017. La veille et le repérage sur des sites majeurs étaient ainsi facilités et permettait d'éviter les doublons dans la base de données de l'outil de sélection. Ainsi, 1395 fiches de sites de précédentes collectes électorales ont été réutilisées en 2017. 892 fiches des élections présidentielles et législatives de 2012, 20 fiches des municipales 2014 et 483 fiches des régionales de 2015 ont été réutilisées. Le processus de récupération de ces fiches devra être amélioré pour les futures collectes car la quantité de sélections et l'absence de modification par lot entre deux collectes différentes rendent de plus en plus fastidieux le transfert de ces fiches.

## 4.2. Volumétrie et répartition statistiques des collections

Les collectes de sites électoraux se sont déroulées sur plus de six mois.
Les collectes représentent un volume de 7 Téraoctets de données pour l'ensemble des collectes, avec respectivement 4.7 Téraoctets pour les collectes mensuelles, 2.2 Téraoctets pour les collectes hebdomadaires, 0.1 pour les collectes pluriquotidiennes.
En nombre de pages, cela représente environ 240 millions de pages, dont 190 millions sont des pages bien collectées en code HTTP 200. La répartition est de 59 millions pour les collectes hebdomadaires, 172 millions pour les collectes mensuelles, et 9 millions pour les collectes quotidiennes. Les pages restantes ne sont pas nécessairement des pages d'erreurs mais des pages de redirection (code 300) ou d'erreurs inhérentes aux sites ou ponctuellement générées par le robot de collecte (code 400).

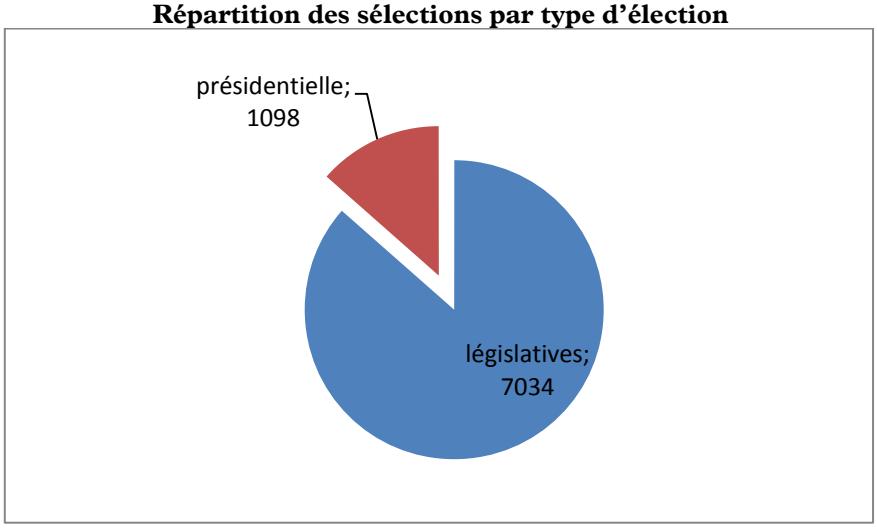
### *4.2.1. Répartition de la sélection*

Au total, **8132 URL de départ** ont été sélectionnées. Inférieur au nombre de sites collectés dans le cadre des élections de 2012, qui représentait environ 11 000 sélections, ce chiffre reste dans la lignée des collectes projets électorales précédentes.

Les sélections ont été réparties entre l'élection présidentielle et les élections législatives : **1098 URL de départ** ont été sélectionnés pour la collecte présidentielle, **7034** pour les collectes des législatives.

*Répartition par élection*

**Répartition des sélections par type d'élection**




Cependant, comme en 2012 ou pour les autres collectes projets, ce chiffre ne suffit pas pour autant à caractériser les collections constituées. En effet, il ne s'agit pas d'autant de sites à proprement parler, mais d'adresses URL qui ont été données au robot comme point de départ de la collecte. Le nombre de sites collectés est forcément supérieur du fait des rebonds de collecte. Cela tient à la nature hypertextuelle des pages et à des nombreuses citations de sites non sélectionnés mais finalement collectés. D'ailleurs, la notion en elle-même de « site » est vaste et potentiellement nébuleuse, allant du site officiel du parti politique au compte Twitter d'un homme politique, en passant par le blog d'un militant.

*Répartition par thématique*

| Type de site | Nombre de sites sélectionnés |
|---|---|
| 0 Sites officiels et institutionnels | 199 |
| 1.1 Sites des candidats en campagne | 3597 |
| 1.2 Sites des formations politiques | 1092 |
| 1.3 Autres organisations de soutien | 416 |
| 2.1 Annuaires, observatoires et analyses | 656 |
| 2.2 Médias traditionnels | 513 |
| 2.3 Associations, syndicats et autres organisations | 315 |
| 2.4 Expressions individuelles et communautaires sur l'Internet | 1233 |

La répartition des sites au sein de la typologie retenue montre que les sites des catégories 1.1 ont fait l'objet d'un plus grand nombre de sélection (44% du total). Il s'agit d'une légère hausse par rapport aux précédentes élections présidentielles et législatives, puisqu'en 2012 cette catégorie de sites représentait 41% du total des sites sélectionnés. La catégorie 2.4, potentiellement la plus vaste car couvrant les expressions du corps social au plus large, est également importante avec environ 15% des sélections. Les sélections se sont concentrées sur les citoyens s'exprimant régulièrement sur le sujet politique. Même s'il n'a pas été possible de collecter les pages publiques Facebook, le taux de sites sélectionnés dans la catégorie 2.4 est légèrement supérieur à celui des collectes présidentielles et législatives de 2012. Cela s'explique notamment par la place toujours plus importante qu'occupent désormais les réseaux sociaux, qui seraient devenus la première source d'information pour près de 40% des jeunes[6]. De fait, ils sont également devenus incontournables pour toutes les formations et personnalités politiques. Chaque parti et chaque candidat disposent de comptes Facebook et Twitter comptant parfois plusieurs millions d'abonnés et cumulant pour certains des millions de vues sur leurs vidéos Youtube.

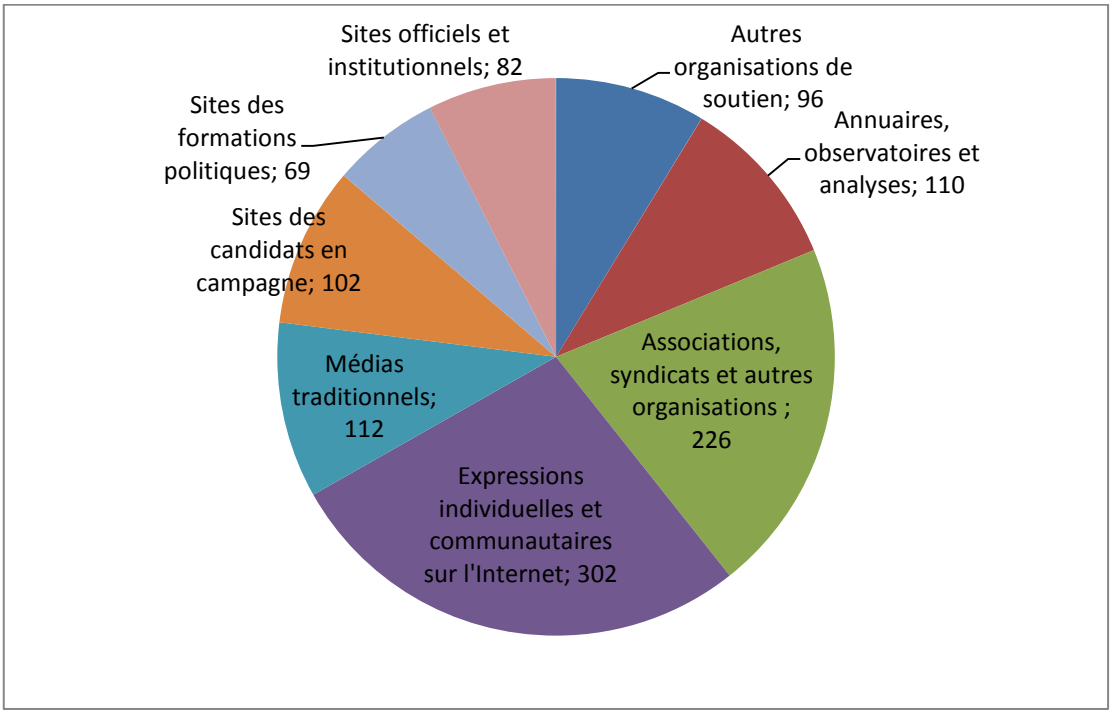
Mais si la campagne électorale occupe toujours plus le terrain du numérique, on notera tout de même une plus faible proportion de blogs dans les sélections de 2017. Alors qu'en 2012 ils représentaient un moyen d'expression citoyenne encore privilégié, ces derniers semblent avoir été délaissés au profit de la plus grande spontanéité des réseaux sociaux et des forums, à l'image des forums du site Jeuxvideos.com, qui fédèrent une communauté importante et très active, surtout chez les 15-25 ans. Autre changement remarquable par rapport aux sélections des campagnes électorales de 2012, la catégorie 1.2 qui représentait 18% des sélections constitue 14% des sites sélectionnés en 2017. Cela peut s'expliquer par le contexte particulier et la singularité de cette campagne électorale. En effet, l'arrivée de nouveaux mouvements politiques (LREM, FI) et l'absence de certains partis plus traditionnels du paysage politique français (EELV, MoDem, PC, UDI) au profit de certaines alliances peuvent en

6 http://www.la-croix.com/Economie/Medias/Barometre-medias-Francais-veulent-information-verifiee-2017-02-02-1200821914

partie expliquer cette baisse dans les sélections de la catégorie 1.2. Enfin, il est possible que les formations politiques et leurs candidats, notamment ceux issus de « petits partis », aient parfois privilégié une présence sur les réseaux sociaux plutôt que la création d'un site, potentiellement éphémère, dédié à la campagne électorale.

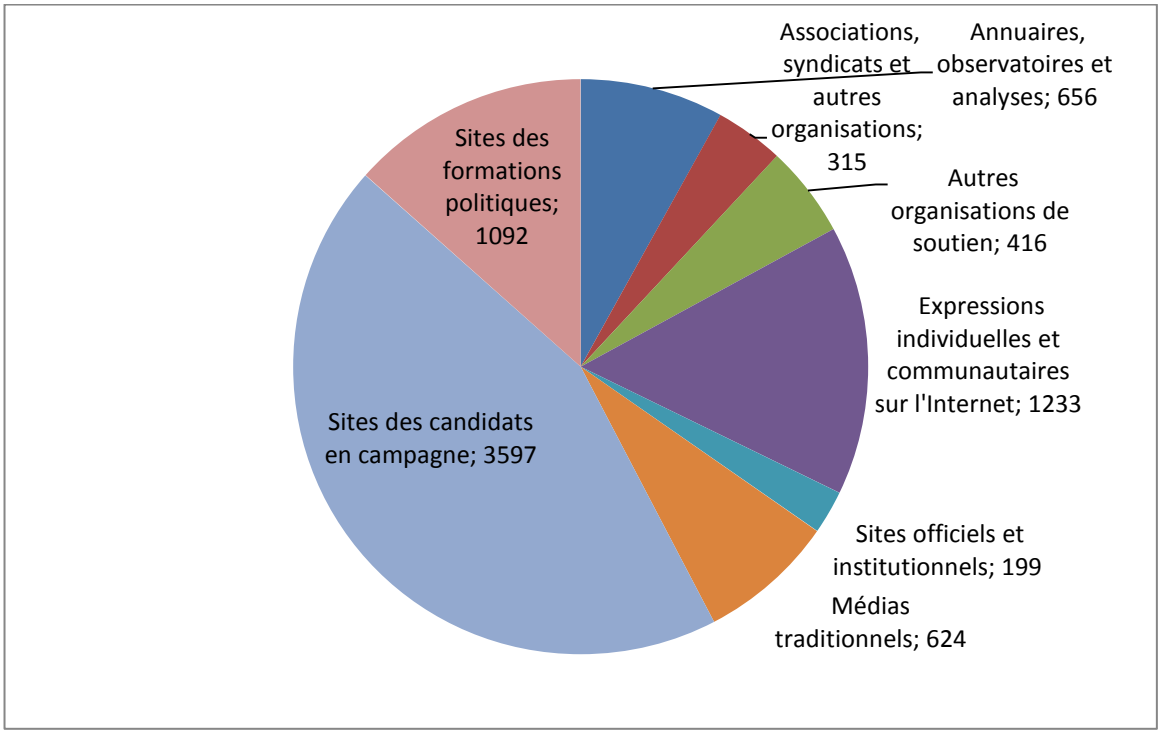
Enfin, la catégorie 2.2 constitue 6% des sites sélectionnés en 2017 contre 9% en 2012. Cette baisse s'explique notamment par le fait que de nombreux titres de presse étant collectés quotidiennement dans le cadre de collectes courantes, il a été décidé de ne pas les collecter, pour la plupart, en doublon.

**Répartition des catégories de la typologie pour la sélection de niveau national**




La répartition des sélections par typologie de site met en évidence que la proportion de sites collectés par catégorie diffère selon les scrutins présidentiels ou législatifs. Ainsi, dans le cadre de la collecte de la présidentielle, les sites les plus sélectionnés par les correspondants sont issus de la catégorie 2.4, suivis par la catégorie 2.3 puis par la catégorie 2.1 et 1.1. Dans les collectes des législatives, c'est au contraire la catégorie 1.1 qui représente le plus grand nombre de sélections, suivie par la catégorie 2.4 puis 1.2. Cela s'explique notamment par un nombre bien plus important de candidats dans le cadre des élections législatives.

**Répartition des catégories de la typologie pour la sélection de niveau national et régional**



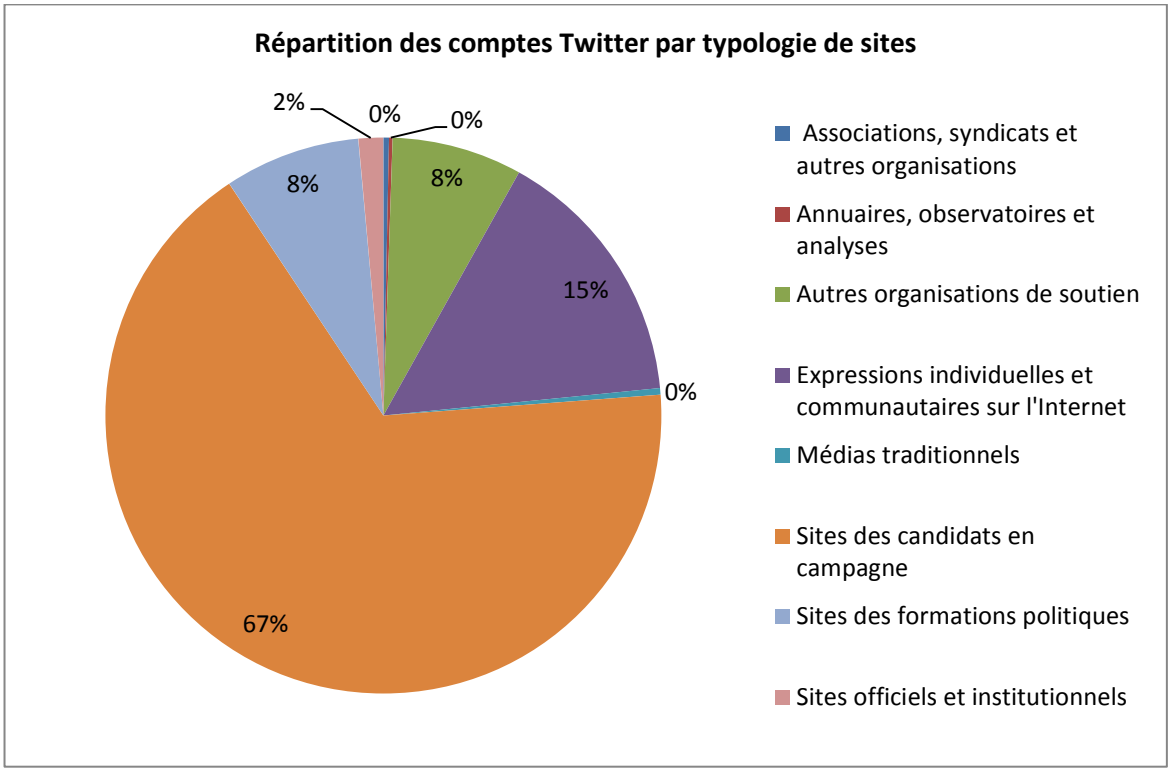

*Réseaux sociaux : focus sur Twitter*

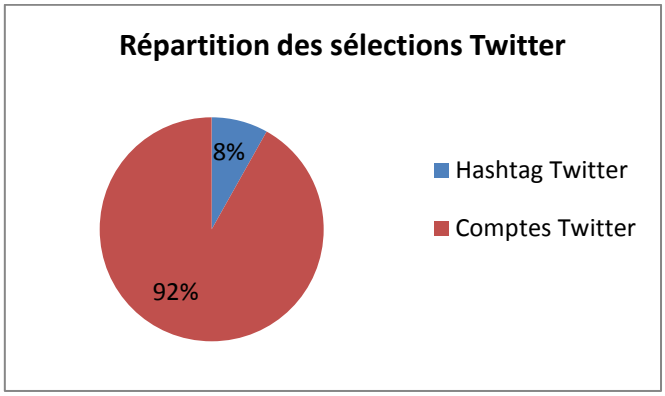
En 2017, l'utilisation massive et quasi systématique des réseaux sociaux à des fins de communication politique est toujours d'actualité. Fin 2016, Twitter inaugurait d'ailleurs un compte dédié à la couverture de l'élection présidentielle française, avec l'objectif de « s'investir dans la vie citoyenne et [de] faciliter la communication directe entre votants et personnel politique »[7]. Le directeur général de Twitter France définissait alors le réseau social comme « un carrefour d'information et de débats dans le cadre des élections à venir »[8]. Ce qui explique en partie la forte proportion, dans les sélections des correspondants, de comptes et de hashtag Twitter. Les agents de la BnF et des BDLI ont ainsi sélectionné 2889 comptes Twitter et 301 hashtags, soit un total de 3190 URL provenant du réseau social, autrement dit près de 40% du total des sélections des collectes du web électoral 2017. En 2012, Twitter représentait environ 10% de l'ensemble des collectes électorales. Concernant les hashtags, il faut noter que ces derniers nécessite une grande réactivité pour leur bonne sélection et gestion dans le cadre des collectes du web électoral. Par définition le hashtag est éphémère et peut devenir, ne serait-ce qu'en quelques jours, obsolète pour la collecte car délaissé par les utilisateurs.

Enfin, la répartition des comptes Twitter par typologie de sites met en évidence que Twitter n'est pas qu'un seul moyen d'expression citoyenne et individuelle : candidats, partis, associations, institutions, médias se sont largement emparés de cet outil de communication 2.0.

7 https://www.challenges.fr/election-presidentielle-2017/comment-twitter-s-invite-dans-l-election-presidentielle-de-2017_442832

8 https://www.challenges.fr/election-presidentielle-2017/comment-twitter-s-invite-dans-l-election-presidentielle-de-2017_442832



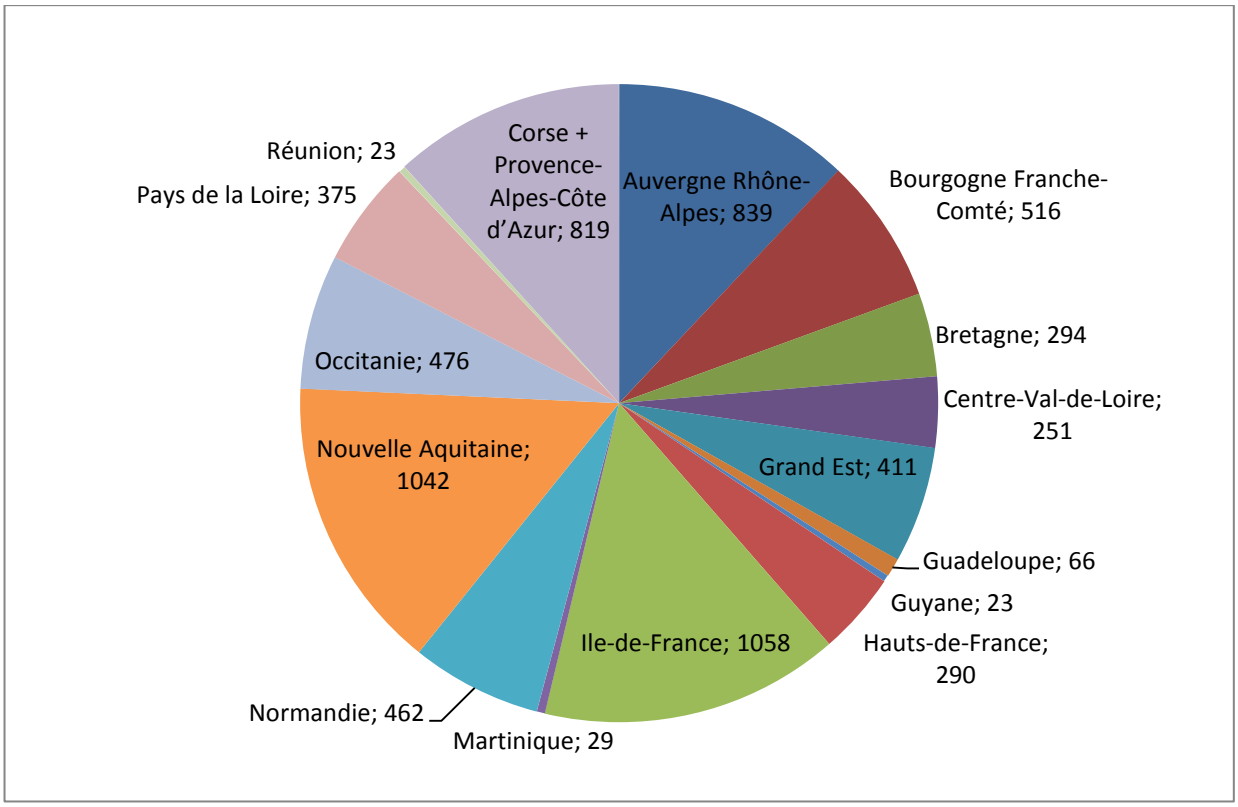

*Législatives : répartition par région*

Le tableau suivant donne un premier aperçu de l'étendu de la campagne en ligne dans chaque région. Les variations entre le nombre de sites sélectionnés selon les régions peuvent notamment s'expliquer par des campagnes en ligne parfois nettement moins actives dans certaines circonscriptions.

| Régions | Nombre de sites sélectionnés |
|---|---|
| Auvergne Rhône-Alpes | 839 |
| Bourgogne Franche-Comté | 516 |
| Bretagne | 294 |
| Centre-Val-de-Loire | 251 |
| Grand Est | 411 |
| Guadeloupe | 66 |
| Guyane | 23 |
| Hauts-de-France | 290 |
| Ile-de-France | 1058 |
| Martinique | 29 |
| Normandie | 462 |
| Nouvelle Aquitaine | 1042 |
| Occitanie | 476 |
| Pays de la Loire | 375 |
| Réunion | 23 |
| Corse + Provence-Alpes-Côte d'Azur | 819 |

*Les campagnes en ligne de Provence-Alpes-Côte d'Azur et de la Corse ont été couvertes par les correspondants de la BnF au sein de la même collecte.

**Répartition des sélections par région**




## 4.3. Signalement et valorisation

Les données des collectes électorales de 2017 ont rejoint des données déjà présentes sur data.gouv.fr, la plate-forme d'Open Data du gouvernement. Un nouveau tableau CSV, dans la continuité et sous la même forme que ceux des précédentes collectes, sera ajouté sur la page correspondante[9]. Il s'agit d'un premier niveau de signalement des collections, à disposition de tous les internautes et étape préalable à la consultation des archives de l'internet dans les salles de recherche de la BnF et des établissements équipés de l'accès à distance.

Les mots-clés saisis dans BCWeb sont indiqués dans les tableaux mis en ligne. En premier lieu utiles pour préciser les délimitations documentaires de la sélection, ils peuvent permettre de retrouver thématiquement un site dans ces listes lorsqu'ils ont été saisis (leur ajout était facultatif). Ils ont été harmonisés par le comité de sélection en plusieurs niveaux et selon un ordre prédéfini[10]:

- le niveau géographique: correspondant au nom de la **nouvelle région**
- l'appartenance partisane de l'URL sélectionnée : lorsqu'il y avait lieu, avec les partis ou unions de partis inscrit en abrégé
- le nom de la personnalité (pour les candidats notamment)
- des mots-clés libres au choix du sélectionneur
- l'historique de l'URL (qui permet de savoir si l'URL a déjà été collectée dans le cadre de collectes électorales antérieures)

En matière de communication sur la collecte du web électoral 2017, plusieurs contenus ont été diffusés :

- un communiqué de presse[11]
- une brève parue sur Biblionautes et sur le site www.bnf.fr

[9] https://www.data.gouv.fr/fr/datasets/collectes-du-web-electoral-par-la-bnf/
[10] Cf. également Annexe 2
[11] http://www.bnf.fr/documents/cp_archive_web_electoral.pdf

- la publication du parcours guidé « Le web électoral de 2010 à 2015» accessible dans les archives de l'Internet. Un article synthétisant ce parcours guidé a également été publié sur le *Blog Lecteurs* de la Bibliothèque nationale de France[12]
- un article dans *Trajectoire,* le magazine interne de la BnF

Par ailleurs, le site *Archimag.com* a publié un article autour de la collecte du web électoral[13], tout comme le site The conversation.com qui a aussi publié un article sur le projet[14]. Côté radio, la chaîne suisse *RTS* a réalisé un reportage sur le dépôt légal du web avec un focus sur la collecte du web électoral. Enfin, côté audiovisuel, *TF1* a réalisé un reportage[15] pour son journal d'informations autour de la collecte électorale 2017 et France 3 Lorraine a réalisé un reportage sur la Bibliothèque de Nancy, évoquant notamment la collecte du web électoral 2017.

## 4.4. Conclusion et perspectives

La forte mobilisation, dans un temps contraint, de tous les acteurs du projet des collectes électorales de 2017 ne peut qu'être mis en avant en conclusion. Elle s'effectue en outre dans la continuité documentaire et organisationnelle des huit précédents projets de collectes électorales au titre du dépôt légal numérique depuis la présidentielle de 2002, et peu de temps après les élections régionales de décembre 2015. Ce travail en coopération, s'appuyant sur les outils communs et les compétences complémentaires de professionnels des bibliothèques, a une nouvelle fois permis une couverture pluraliste et équilibrée d'une majeure partie des publications en ligne et né-numériques relatives à ce qui demeure les temps forts médiatiques de la vie politique française : les élections présidentielle et législatives. Techniquement, le nouveau processus de collectes des comptes et mots-dièse Twitter vient compléter des captures d'une très bonne qualité pour tous les sites web classiques. Il demeure quelques obstacles à la récupération optimale de contenus importants, comme les pages publiques du réseau social Facebook ou les publications vidéo des candidats, mais des solutions, telles que le recours en 2017 à la prestation externe pour une bonne collecte des vidéos, sont de plus en plus envisageables et envisagés. Après avoir mené à bien, dans un calendrier resserré, les missions primordiales du dépôt légal- collecter et préserver- la mission tout aussi essentielle de la mise en accès de ces collections pour les chercheurs et citoyens est l'enjeu qui se pose aujourd'hui et pour les mois, années à venir. La publication des listes d'adresses URL de sites sur la plateforme ouverte des données publiques françaises va en ce sens https://www.data.gouv.fr/fr/datasets/collectes-du-web-electoral-par-la-bnf/ tout comme la multiplication des lieux de consultations des archives de l'internet (voir la carte des établissements proposant l'accès en leurs murs). Et tout cela en attendant les innovations à venir en matière d'exploration de ces collections patrimoniales représentant un gisement précieux de sources pour l'histoire et l'exercice de la vie démocratique et de la citoyenneté en France à l'ère d'internet.

[12] http://blog.bnf.fr/lecteurs/index.php/2017/05/a-la-decouverte-du-web-electoral-de-2010-a-2015/
[13] http://www.archimag.com/archives-patrimoine/2017/03/15/presidentielles-2017-bnf-archive-web-electoral
[14] http://theconversation.com/archives-comment-le-web-devient-patrimoine-76487
[15] http://www.lci.fr/societe/quand-la-bibliotheque-nationale-de-france-archive-le-net-pour-les-generations-futures-2045544.html

## 5. Annexe 1 – Liste normalisée des noms de partis

- Extrême gauche ou gauche antilibérale

Nouveau parti anticapitaliste (NPA)
Lutte ouvrière (LO)
Front de gauche (FG)
Parti de gauche (PG)
Parti communiste français (PCF)
Parti ouvrier indépendant démocratique (POID)
La France insoumise (FI)

- Gauche :

Parti Socialiste (PS)
Europe Ecologie Les Verts (EELV)
Union des démocrates et des écologistes (UDE)
Parti Radical de Gauche (PRG)
Mouvement 100% (M100%)
Mouvement des progressistes (MDP)
Mouvement écologiste indépendant (MEI)
Mouvement républicain et citoyen (MRC)
Nouvelle donne (ND)

- Centre :

Mouvement Démocrate (MoDem)
Union des démocrates et indépendants (UDI)
Parti Chrétien-Démocrate (PCD)
En Marche ! (EM!)
La république en marche (LREM)*
Parti fédéraliste européen (PFE)

- Droite :

Les Républicains (LR)
Debout La France (DLF)
Union populaire républicaine (UPR)

- Extrême droite

Front national (FN)

- Autres :

Solidarité et progrès (SP)
Résistons ! (RE)

Pour les partis ou mouvements qui n'ont pas une représentation nationale, mettre en mot-clé :

- **DUD** pour divers droite ou union de droite
- **DUG** pour divers gauche ou union de gauche
- **DVE** pour divers écologistes
- **DVC** pour divers centristes
- **DED** pour divers extrême-droite
- **PL** pour partis à ancrage local (exemple : UL (Unser Land) ou AL (Alsace d'abord)
- **DUT** pour divers thématiques ou union thématique pour un parti sur un thème, comme « Mouvement 100% » ou « Chasse, pêche, nature et traditions »
- « **autre** » pour les mouvements politiques indéterminés.

Préciser en note de contenu la nomenclature exacte, en mettant le libellé complet, si besoin.

* Après l'élection présidentielle, le mouvement a changé de nom : "En Marche !" est devenu "La république en marche". On utilisera donc cette dernière formulation dans le cadre des élections législatives de 2017.

## 6. Annexe 2 – Calendrier des dates de collecte

Collecte Mensuelle

| | |
|---|---|
| Date de lancement de la 1ere collecte | **Jeudi 1er janvier 2017 (pour le premier tour des primaires)** |
| Date de lancement de la 2ere collecte | **Mercredi 1er février (pour le deuxième tour des primaires)** |
| Date de lancement de la 3e collecte | **Mercredi 1er mars** |
| Date de lancement de la 4e collecte | **Mardi 4 avril** |
| Date de lancement de la 5e collecte | **jeudi 27 avril (entre les deux tours de la présidentielle)** |
| Date de lancement de la 6e collecte | **Lundi 15 mai** |
| Date de lancement de la 7e collecte | **Jeudi 1er juin** |
| Date de lancement de la 8e collecte | **Mardi 13 juin** |
| Date de lancement de la 9e collecte | **Jeudi 22 juin** |
| Date de lancement de la dernière collecte | **Mercredi 12 juillet 2017** |

Collecte Hebdomadaire

| | |
|---|---|
| Date de lancement de la collecte | **Lundi 13 février 2017 - 14h** |
| Date de fin de la collecte | **Mercredi 12 juillet 2017** |

Collecte pluriquotidienne (deux collectes par jour à 11h et 23h)

| | |
|---|---|
| Date de lancement de la 1ere collecte | **Lundi 20 mars 2017** |
| Date de lancement de la dernière collecte | **vendredi 30 juin 2017** |